Институт философии РАН
Кафедра истории и философии науки

Федеральное государственное бюджетное учреждение науки
Вычислительный центр
им. А.А. Дородницына
Российской академии наук

РЕФЕРАТ
по истории и философии науки

История непараметрических байесовских методов в машинном обучении

Специальность
05.13.11
Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Выполнил:
Бартунов Сергей Олегович, аспирант
Научный руководитель:
Сенько Олег Валентинович, д.ф.-м.н.

Москва
2013

Оглавление

Введение 2
Первая работа 6
Развитие непараметрических случайных процессов 7
История приложений непараметрических байесовских методов 13
Эволюция методов вывода для непараметрических моделей 15
Заключение 18
Список литературы 18

Введение

Машинное обучение - чрезвычайно молодой раздел науки, по некоторому мнению сформировавшийся окончательно на стыке множества различных дисциплин в начале 70-х годов прошлого века, когда наши соотечественники Вапник и Червонекис математически обосновали принцип минимизации эмпирического риска[1].

Одним из принципиально важных следствий этого результата стала возможность формулировать задачи машинного обучения как задачи численной оптимизации, а также формирование определенной методологии исследования алгоритмов обучения. Кратко, эта методология заключается в том, что исследователь должен собрать выборку примеров, для каждого из которых известен правильный ответ, после чего эта выборка делится на две (в простейшем случае) части, на одной из которых алгоритм обучается, а на другой тестируется. Помимо этого исследователь должен зафиксировать некоторую функцию потерь, которая для каждого примера и ответа алгоритма на этом примере определяет, насколько сильно ошибся алгоритм. Далее процесс обучения можно представить как выбор такого алгоритма из некоторого семейства (например, заданного параметрически), который ошибается как можно меньше в среднем по обучающей выборке.

Доказанный Вапником и Червонекисом факт позволял надеяться на то, что при достаточно большом объеме обучающей выборки, эмпирический риск, то есть, средняя ошибка по обучающей выборке, не будет сильно отклоняться от вероятности ошибки на генеральной совокупности примеров. После этого открытия было предложено множество различных функций

[1] (Vapnik & Chervonenkis, 1971)

потерь, учитывающих специфику определенных задач, а также проанализированы те или иные их свойства (выпуклость, монотонность, непрерывность и т.д.), специальных методов оптимизации, работающих эффективнее методов общего назначения, а также выделены интересные семейства алгоритмов, в частности, линейные модели, деревья решений, нейросети и т.д.

Тем не менее, наивная минимизация эмпирического риска в реальных условиях при решении конкретных прикладных задач далеко не всегда приводила к приемлемому результату. Несмотря на то, что в результате обучения выбирался оптимальный с точки зрения эмпирического риска алгоритм, его эффективность на тестовой выборке могла быть произвольно низкой при малых ее размерах относительно размера рассматриваемого семейства алгоритмов (в статистическом смысле), а также в ряде других случаев, нередко встречающихся на практике.

Теория Вапника-Червоникиса (а позднее и другие теории) формально предоставляла оценку необходимой размерности обучающей выборки, которая гарантировала бы сходимость эмпирического риска, однако, эта оценка была сильно завышенной и не могла достигаться на практике, кроме того, хорошие результаты обучения достигались зачастую и на малых выборках.

Другой серьезной проблемой, помимо недостаточного размера выборки, было отсутствие понимания, как именно определять и в дальнейшем работать с нужными семействами алгоритмов. Как правило, эти семейства описывались параметрически, например, в случае задачи бинарной классификации, когда выборка могла быть достаточно точно разделена на два класса гиперплоскостью в соответствующем пространстве признаков, множество алгоритмов могло быть задано как множество всех гиперплоскостей в пространстве заданной размерности. На ряде задач такой подход давал хорошие результаты, однако, казалось бы, незначительное усложнение постановки задачи наглядно демонстрировало ограниченность параметрической методологии.

В качестве важного примера можно рассмотреть классическую задачу кластеризации множества точек в эвклидовом пространстве, то есть, разбиения этого множества на некоторые группы точек, так называемые кластеры, которые имеют некоторый устойчивый смысл, то есть, чтобы при кластеризации другой выборки, взятой из того же распределения, найденные группы по-прежнему имели бы смысл. Примечательно, что до сих пор в научном сообществе нет единого мнения относительно того, что считать хорошей кластеризацией, и критерии качества (иначе говоря, соответствующие функции потерь) выбираются отдельно для каждой задачи.

В одной из классической формулировок, в роли кластеров выступают многомерные нормальные распределения, а принадлежность каждой точки некоторому кластеру выражается при помощи соответствующих вероятностей.

Данная постановка задачи подробно исследуется в рамках математической статистики, где она получила название задачи разделения смеси распределений. Традиционно, она решается при помощи метода максимального правдоподобия, предложенного Фишером, в предположении об известном числе компонент смеси. Однако, если число компонент смеси сделать переменным, то максимум функции правдоподобия достигается, когда каждая точка образует свой собственный кластер, состоящий из себя самой!

Таким образом, появились понятия “сложности модели” и “структурного параметра”, точнее, они, как и многие другие понятия, были адаптированы из математической статистики. В последнем примере в роли сложности модели выступало число кластеров или распределений, которые, в свою очередь, являлись структурными параметрами. Другими примерами структурных параметров являются скрытые слои в многослойных перцептронах или факторы в латентных факторных моделях. Вскоре стало очевидным, что при оптимизации эмпирического риска (или при решении любой другой задачи оптимизации в обучении) необходимо контролировать сложность модели. Один способов такого контроля получил название “регуляризация” и зачастую заключался в том, что к оптимизируемой функции добавлялся некоторый член, штрафующий результирующую модель за ее излишнюю сложность. Кроме того, появилось несколько теорий, позволяющих объяснить феномен “переобучения”, а также избежать его, в частности, информационный критерий Акаике, связывающий число параметров в модели и результирующее значение функции правдоподобия на обучающей выборке. Позднее Гидеон Шварц разработал байесовский информационный критерий, в котором использовал те же идеи, что и Акаике, но исходил из принципов байесовской статистики. Оба этих критерия имеют в качестве аналогии широко известный принцип Оккама, согласно которому из двух моделей, показывающих одинаковый результат на обучающей выборке, стоит предпочесть ту, которая использовала меньшее количество параметров.

В виду того, что в основе теории машинного обучения изначально лежали результаты из множества других научных дисциплин, на протяжении всего времени у исследователей возникали идеи по применению тех или иных методов из родственных наук для решения ключевых проблем вроде выбора модели. Исключением не стала и байесовская статистика, которая с

середины 90-х годов и по настоящий момент переживает ренессанс именно в связи с тем, что она нашла свое применение в бурно развивающимся машинном обучении.

Причины популярности байесовского подхода в машинном обучении заключаются вовсе не в интерпретации понятия вероятности, как меры уверенности (в противовес частотному определению), вопреки распространенному убеждению, а скорее в том, что байесовская парадигма предполагает конструирование моделей с использованием определенного математического языка, владение которым позволяет быстро комбинировать мелкие составляющие в единую модель.

Все рассматриваемые объекты, участвующие в модели, определяются как случайные величины, взаимодействующие друг с другом согласно соответствующим факторам, как правило, имеющим вероятностный смысл. Часть из этих величин может считаться “наблюдаемыми”, часто признаки обучающих примеров выступают в этой роли, другая часть называется “скрытыми” и может отвечать за настраиваемые параметры, остальные величины называются “гиперпараметрами” и фиксируются как константы. Модульность построения вероятностных моделей в данном случае заключается в том, что гиперпараметры, воспринимаемые как константы, могут быть заменены на случайные величины, определенные в другой модели, и, например, стать настраиваемыми параметрами после объединения моделей.

Помимо этого, в отличие от самых ранних алгоритмов обучения, в которых настройка параметров заключалась в их оптимизации, байесовский подход предполагает так называемый “байесовский вывод” в отношении параметров, то есть, получение условного распределения на значения скрытых переменных при условии наблюдаемых. Поскольку на практике обучающая выборка часто зашумлена или имеет перекосы, то встроенные механизмы моделирования неопределенности в принятии решений оказались очень кстати.

Стали формулироваться вероятностные аналоги ранее известных невероятностных методов, в которых последние возникали как частные случаи, соответствующие оценке максимального правдоподобия в соответствующих распределениях; в качестве примера можно привести байесовский метод главных компонент[2], предложенный одним из ведущих специалистов по байесовским методам Кристофером Бишопом.

Кроме того, было продемонстрировано, что во многих случаях регуляризация невероятностных моделей соответствует определенным

[2] (Bishop, 1999)

априорным распределениям в вероятностных[3]. Это в свою очередь заложило основу для непараметрического байесовского подхода к задаче выбора модели. Вместо того, чтобы обучать несколько алгоритмов с различными структурными параметрами, а потом выбирать из них наилучший при помощи, например, критерия Акаике[4], было предложено настраивать только одну модель со специальным априорным распределением на структурные параметры.

Этот подход, в частности, привнес весьма интересные свойства в модели, позволяя структуре модели меняться по мере поступления новых обучающих данных, без перезапуска процедуры обучения, однако потребовал некоторого изменения методологии обучения и образа мышления желающих применять непараметрические методы. Об этом подходе далее и пойдет речь.

Грубо говоря, семейство непараметрических методов можно разделить на две группы: связанные с гауссовскими процессами и связанные с процессом Дирихле. Эти группы развивались независимо и не пересекались по той причине, что методы из каждой группы решали свои задачи. Несмотря на то, что формально гауссовский процесс является непараметрическим распределением наравне с процессом Дирихле, особенно в последнее время под непараметрическими процессами понимают именно вторую группу процессов, и из данного понятия почти полностью вытеснены гауссовские процессы. По этой причине в данном реферате будут рассмотрены только методы второй группы.

Первая работа

В 2000 году на одной из главных международных конференций по машинному обучению Neural Information Processing Systems (NIPS) была опубликована статья Карла Эдварда Расмуссена “The Infinite Gaussian Mixture Model”[5]. Это была одна из первых статей, демонстрирующая использование непараметрических байесовских методов в машинном обучении.

В статье рассматривалась классическая задача разделения смеси и главным образом внимание уделялось априорному распределению на принадлежность компонентам смеси. Традиционно для этого применялось распределение Дирихле заданной размерности. В своей работе Расмуссен рассмотрел один из методов байесовского вывода, а именно схему генерации

[3] (Bishop & Nasrabadi, 2006)
[4] (Akaike, 1974)
[5] (Rasmussen, 2000)

выборки по Гиббсу (далее – схема Гиббса) для модели с распределением Дирихле, после чего перешел к пределу относительно числа кластеров. Оказалось, что предел существует и имеет очень простой и понятный смысл: вероятность принадлежности каждой точки к любому из возможных кластеров пропорциональна числу точек, уже отнесенных к этому кластеру. Кроме того, точка имеет возможность образовать новый кластер с вероятностью, пропорциональной параметру концентрации симметричного бесконечномерного распределения Дирихле. Несмотря на всю простоту этого результата, он имел очень серьезное значение: ранее не предполагалось, что образование новых кластеров может быть промоделировано априорным распределением.

Полученная Расмуссеном конструкция ранее уже была открыта французским математиком Давидом Альдоусом[6] и позднее была названа процессом китайского ресторана[7]. Примечательно, что этот неожиданно возникший в работе Расмуссена случайный процесс, как было показано позднее, является одним из представлений процесса Дирихле, который сам Расмуссен цитирует в своей статье, но, судя по всему, не видит связи между двумя этими процессами.

Эта работа Расмуссена не предложила саму концепцию использования непараметрических байесовских методов, которая была известна ранее, но зато продемонстрировала крайней простой практический метод вывода, что подстегнуло интерес научного сообщества к непараметрическим байесовским методам.

В дальнейшем область развивалась по следующим направлениям: получение новых моделей из старых, путем использования непараметрических распределений вместо параметрических, и исследование свойств новых моделей, разработка более эффективных методов байесовского вывода для этих моделей, а также поиск новых непараметрических случайных процессов.

Развитие непараметрических случайных процессов

[6] (Aldous, 1985)

[7] Такое название этот случайный процесс получил из-за наглядной метафоры поведения посетителя в китайском ресторане, который садится за любой из возможных столов с вероятностью, пропорциональной числу сидящих за ним людей. В дальнейшем другие случайные процессы стали также называть в честь заведений общественного питания.

Первым из непараметрических случайных процессов был процесс Дирихле[8], предложенный Томасом Фергуссоном в 1973 г. В своей работе Фергуссон предложил сразу два эквивалентных определения процесса Дирихле, первое из которых было схоже с одним из определений гауссовского процесса и рассматривало свойство любой выборки из процесса Дирихле быть распределенной согласно распределению Дирихле с определенными параметрами. Второе же основывалось на представлении процесса с точки зрения понятия *случайной меры*.

Процесс Дирихле не ограничился двумя определениями. Будучи интересным математическим объектом, полезным для многих практических задач, он привлек к себе внимание научного сообщества, в результате чего для него были получены различные эквивалентные определения или, иначе говоря, представления, оперирующие на первый взгляд не имеющими ничего общего понятиями, но на самом деле описывающих один и тот же случайный процесс.

Первым разработанным представлением была конструкция на основе урны Пойа[9], предложенная в 1984 г. Фредом Хоппе, в честь которого ее иногда называют. В данном случайном процессе моделировалось распределение над разбиениями множества натуральных чисел, путем обобщения модели урны Пойа. Спустя год было предложено другое распределение, впоследствии получившее название процесса китайского ресторана[10], о котором уже было сказано выше. В обоих этих работах не упоминалось связи с процессом Дирихле, хотя ее можно было проследить, так как модель урны Пойа тесно связана с распределением Дирихле, а процесс китайского ресторана, в свою очередь, полностью эквивалентен урне Хоппе. Тем не менее, оба этих распределения строго говоря не являлись эквивалентными процессу Дирихле, поскольку их носителем являлось другое множество. В дальнейшем процесс китайского ресторана стал часто применяться для определения непараметрических моделей, для которых вывод осуществлялся при помощи схемы Гиббса, поскольку он был наиболее удобен для этого.

В 1994 г. было также открыто другое, крайне важное представление процесса Дирихле, названное процессом разлома палки[11], которое определяло процесс Дирихле *конструктивно* и позволяло, например, аналитически получить вероятность, с которой должен повторяться каждый

[8] (Ferguson, 1973)
[9] (Hoppe, 1984)
[10] (Aldous, 1985)
[11] (Sethuraman, 1994)

из уникальных элементов выборки, сгенерированной из данного случайного процесса. Впоследствии, процесс разлома палки оказался наиболее удобен для вариационного вывода, в частности, по той причине, что он, как и распределение Дирихле, был сопряженным априорным распределением к мультиномиальному. Стоит тем не менее упомянуть, что при этом использовался так называемый ограниченный процесс разлома палки (англ. truncated stick breaking process), в котором максимальное число структурных параметров искусственно ограничивается сверху.

Использование данного процесса, как правило, не создавало трудностей для непосредственного вывода параметров модели, но оно представляло серьезную идеологическую и методологическую проблему. Абсолютное большинство работ, в которых использовался процесс Дирихле и его представление именно в виде процесса разлома палки (с целью вариационного вывода), формально работали с непараметрическими моделями, однако, по сути эти модели были аналогичны параметрическим! Кроме того, это ограничение при приближенном выводе занижало дисперсию апостериорного распределения[12].

В 1997 г. было предложен процесс Питмана-Йора[13], который в некотором смысле обобщал процесс Дирихле и позволял получать другой характер угасания вероятностей повторения уникальных элементов из базовой меры. В частности, удавалось добиться степенного закона убывания этих вероятностей, которому часто отводится особое место в описании различных процессов.

Еще один важный непараметрический процесс, названный процессом индийского буфета[14], был предложен в 2005 году как распределение над классами эквивалентности бинарных матриц бесконечной размерности. Грубо говоря, если в процессе китайского ресторана моделировалась ситуация, в которой каждый объект мог принадлежать одному и только одному классу (разбиению), то в процессе индийского буфета объекты могли принадлежать произвольному числу классов, число которых зависело логарифмически от числа объектов. Таким образом, процесс индийского буфета отлично подходил для моделирования моделей латентных свойств (англ. Latent Feature Models).

Чуть позже, в 2007 году, была показана важная связь процесса индийского буфета с бета процессом[15], а также получено его конструктивное

[12] (Wang & Blei, 2012)
[13] (Pitman & Yor, 1997)
[14] (Griffiths & Ghahramani, 2005)
[15] (Thibaux & Jordan, 2007)

представление, аналогичное процессу разлома палки для процесса Дирихле[16]. Результаты, полученные в этих работах, позволили унифицировать теоретические методы работы с рядом непараметрических процессов (включая процесс Дирихле, процесс Питмана-Йора и процесс бета-Бернулли) и глубже понять их свойства как случайных мер[17].

Среди прочих интересных непараметрических распределений стоит также отметить иерархический процесс Дирихле (ИПД), предложенный в 2005 году группой исследователей[18]. Данный процесс представлял из себя двухуровневый генеративный процесс, на первом уровне которого при помощи корневого процесса Дирихле выбирался процесс Дирихле второго уровня, из которого уже генерировался объект. В отличие от обычного процесса Дирихле, ИПД позволял моделировать смесь распределений не только в рамках одной выборки, но и одновременно для нескольких выборок, которые по предположению имели общие параметры компонентов смеси. Это позволяло применять непараметрические методы в ситуациях, когда в распоряжении имеется несколько выборок, полученных, вероятно, в несколько различных условиях, однако имеющих частично *общую природу*, и при этом использовать данные каждой выборки для оценивания общих параметров, одновременно выполняя кластеризацию как внутри каждой выборки независимо, так и кластеризацию кластеров, полученных для каждой выборки. Помимо этого, ИПД можно было использовать и для ситуации, когда в распоряжении имеется только одна выборка, но требуется промоделировать подобный эффект разнородности данных в ней.

Подобное применение ИПД существенно улучшало результаты в задаче разделение смеси, и поэтому во многих непараметрических моделях использовался непосредственно ИПД, вместо обычного процесса Дирихле.

Тем не менее, ИПД был не единственным непараметрическим распределением, призванным смоделировать имеющиеся в данных иерархические зависимости. Был также предложен вложенный процесс китайского ресторана[19] (одновременно с данной работой другой коллектив ученых опубликовал эквивалентный вложенный процесс Дирихле[20]), который позволял не просто моделировать неопределенное число кластеров в выборке, но также их структуру в виде дерева. ВПКР был предложен не как абстрактный статистический метод, а в качестве средства для построения

[16] (Teh, Gorur, & Ghahramani, 2007)
[17] (Broderick, Jordan, & Pitman, 2012)
[18] (Teh, Jordan, Beal, & Blei, 2006)
[19] (Blei, Griffiths, & Jordan, 2010)
[20] (Rodriguez, Dunson, & Gelfand, 2008)

более детальных *тематических моделей,* часто выбираемых в качестве приложения исследователями в области непараметрических методов, и для данной задачи показал крайне хорошие результаты.

До сих пор все рассмотренные случайные процессы предполагали бесконечную симметрическую зависимость моделируемых данных (англ. infinite exchangeability), крайне важное и часто используемое свойство в байесовской статистике, означающее, что произвольная (бесконечная) перестановка данных не меняет их совместного распределения. Важным частным случаем этого свойства является независимость и одинаковое распределение случайных величин, которое еще более часто встречается в математической статистике, однако, является более сильным свойством.

Бесконечная симметрическая зависимость, грубо говоря, предполагает, что порядок, в котором рассматриваются данные, не имеет значения. Это предположение является адекватным для ряда задач, однако, во многих случаях данные имеют внутри себя различные зависимости, несовместимые с этим предположением. Так, например, при анализе изображений или временных рядов порядок рассмотрения соответствующих объектов (точек изображения или элементов ряда) может быть определяющим для результатов. Кроме того, согласно теореме Де Финетти[21], для симметрически зависимых данных возможно разложить совместное распределение на априорное распределение на некий *скрытый параметр,* вообще говоря, бесконечномерный и правдоподобие данных, относительно значения этого параметра, благодаря чему создается теоретическая обоснование для существования непараметрических распределений.

Однако, поскольку различные приложения, в которых порядок данных был важен, требовали инструментов статистического анализа, стали появляться различные расширения классических непараметрических методов, которые не только позволяли учесть существующие зависимости в данных и за счет этого повысить точность получаемых результатов, но и избавляли от необходимости использования более ранних методов в сочетании с ненадежными эвристиками. Например, в работах по моделированию временных зависимостей в данных нередко можно было встретить эвристики, использующие различные вариации метода скользящего окна, в которых для получения качественных результатов требовалось тщательно настраивать размер и другие параметры окна.

Первым из подобных расширений был зависимый процесс Дирихле[22], опубликованный в 1999 году МакИкерном, а так же ряд других процессов,

[21] (de Finetti, 1931)
[22] (MacEachern, 1999)

разработанных позднее, которые объединяло то, что они обобщали обычный процесс Дирихле, делая его зависимым от какой-либо величины (например, времени или координаты в пространстве), каждое отдельное значение которой как бы порождало свой независимый процесс Дирихле.

Другим расширением стал недавно предложенный учеными Блеем и Фрейзером метрически-обусловленный процесс китайского ресторана (англ. distance-dependent Chinese Restaurant Process)[23], который в свою очередь обобщал процесс китайского ресторана, однако не был обобщением процесса Дирихле с точки зрения случайной меры. Если зависимый процесс Дирихле был построен на основе конструктивного представления процесса Дирихле, то моПКР определялся в пространстве разбиений последовательности натуральных чисел и использовал для моделирования зависимостей внутри данных априорно известные попарные расстояния для всех рассматриваемых объектов, которые могли иметь произвольную природу и не обязаны были образовывать настоящее метрическое пространство. Данные расстояния преобразовывались в парные вероятности "сцепления" двух объектов в один кластер при помощи специальных функций угасания.

Несмотря на то, что, на первый взгляд, оба этих семейства распределений предоставляли одинаковые возможности для статистического моделирования не-симметрически зависимых данных, их различие выражалось в достаточно тонком свойстве *маргинальной инвариантности* (англ. marginal invariance), которое заключалось в том, что отсутствие информации о каком-либо объекте в данных не меняло процесс вывода параметров для всех остальной выборки, как если бы отсутствие наблюдаемой информации было эквивалентным отсутствию самого объекта. Данное свойство присуще как обычному процессу Дирихле, так и зависимому процессу Дирихле и аналогичным процессам, но не выполняется для моМКР, что делает его интересным непараметрическим распределением, однако не позволяет с ним работать, как со случайной мерой в общем случае.

Увеличение популярности непараметрических методов вызвало ощущение, что они являются в некотором роде панацеей и про параметрические аналоги стоит забыть. Однако, было замечено, что в случае, когда выборка была получена из конечной смеси на основе распределения Дирихле с известным числом компонент, процесс Дирихле, будучи бесконечномерным обобщением распределения Дирихле, не способен восстановить верное число компонент смеси даже в достаточно простых случаях. В 2013 году Джеффри Миллером и Мэтью Харрисоном была опубликована статья с говорящим названием "A simple example of Dirichlet

[23] (Blei & Frazier, 2011)

process mixture inconsistency for the number of components"[24], наглядно демонстрирующая этот эффект. Не исключено, что эта проблема будет одной из главных, стоящих перед исследователями в области непараметрических байесовских методов в ближайшие годы.

История приложений непараметрических байесовских методов

Поскольку проблема выбора модели и автоматического подбора структурных параметров актуальна для абсолютно всех задач машинного обучения и смежных областей, непараметрические методы в настоящий момент используются повсеместно, вследствие чего невозможно изложить полную историю их приложений. Более того, на волне популярности непараметрических методов появилось множество работ, в которых соответствующие параметрические априорные распределения были заменены на непараметрические без каких-либо интересных следствий, кроме возможности работы с неопределенным числом параметров, и потому едва ли представляющих интерес с исторической точки зрения. Тем не менее, можно выделить наиболее важные работы, популяризовавшие применение непараметрических байесовских методов.

Поскольку наиболее очевидной и одной из классических задач, в которых остро встает вопрос выбора модели, является задача кластеризации (а также ее варианты), как уже было замечено выше, первой работой, в которой были применены методы непараметрической статистики, была работа Расмуссена[25], в которой была представлена "бесконечная" модель смеси нормальных распределений. Следом за ней, в 2001 г. коллектив авторов, в который также входил Расмуссен, разработал другую "бесконечную" модель смеси распределений, но на сей раз сформулированную в виде генеративного процесса, протяженного во времени, в рамках которого данные генерируются из сменяющих друг-друга (переключающихся) компонентов смеси, в некотором смысле аналогично тому, как это происходит в классической скрытой Марковской модели, но с возможностью автоматически определять появление новых состояний (компонент смеси) по мере поступления новых данных[26]. Как и в предыдущей работе Расмуссена, было использовано дискретное

[24] (Miller & Harrison, 2013)
[25] (Rasmussen, 2000)
[26] (Beal, Ghahramani, & Rasmussen, 2001)

представление процесса Дирихле (процесс китайского ресторана), а вывод осуществлялся при помощи схемы Гиббса.

После того, как в 2006 году Блеи и Джордан разработали алгоритм вариационного вывода для процесса разлома палки[27], модели, использующие процесс Дирихле, стали существенно чаще определять именно с использованием данного представления с целью дальнейшего применения вариационного вывода.

Помимо задачи разделения смеси, другой областью-заказчиком для непараметрических методов стало как раз набиравшая в 2000-х годах *тематическое моделирование*. Во многом это произошло по причине того, что коллективы ученых, занимавшихся непараметрической статистикой и тематическими моделями существенно пересекались. С другой стороны, в то время как задача разделения смеси в классическом ее понимании (получение параметров гауссиан и принадлежностей точек эти гауссианам) была достаточно "сухой", тематические модели, будучи методами анализа текстов, имели весьма наглядные приложения и были тем более актуальны в связи с бурным развитием области информационного поиска и интернета в целом. Было опубликовано немало работ, в которых одновременно исследовались и непараметрические распределения и их приложения к тематическому моделированию, так, например, был предложен вложенный процесс китайского ресторана[28], а также различные методы вывода[29] для непараметрических моделей.

Во второй половине 2000х годов непараметрические методы стали применяться и за пределами традиционных для байесовской статистики задач, например, для анализа социальных сетей[30], в биоинформатике[31] или в социальных исследованиях (особенно интересна статья, где при их помощи моделировалась статистика попыток суицида[32]).

Также непараметрические методы нашли свое применение в компьютерном зрении, области, казалось бы, безнадежно "захваченной" другими парадигмами машинного обучения и даже байесовскими методами, среди которых в данной области традиционно были популярны ненаправленные графические модели, некоторым образом

[27] (Blei & Jordan, 2006)
[28] (Rodriguez, Dunson, & Gelfand, 2008)
[29] (Wang & Blei, 2012)
[30] (Xu, Tresp, Yu, & Yu, 2008)
[31] (Dunson, 2010)
[32] (Ruiz, Valera, Blanco, & Perez-Cruz, 2012)

противопоставляемые[33] направленным, в которых обычно используются непараметрические распределения. Одним из существенных преимуществ подобных моделей по сравнению, например, с моделями на основе случайных марковских полей, была возможность сегментации изображений без предварительного обучения на размеченной обучающей выборке. Так, для сегментации изображений были успешно применены процесс Питмана-Йора в 2008г.[34], а в последствие и метрически-обусловленный процесс китайского ресторана[35].

Другой прикладной областью, в которой непараметрические методы получили широкое распространение, стала обработка естественного языка. На основе процессов Дирихле[36], Питмана-Йора[37] и моПКР[38] были сформулированы языковые модели. Иерархический процесс Дирихле был также применен для автоматического получения грамматики языка[39].

С открытием в 2007 году процесса индийского буфета стало также появляться огромное число моделей латентных свойств, включающих в себя важную байесовскую модель матричного разложения[40] и другие методы факторного анализа[41].

Таким образом, начался и до сих пор не завершился процесс “непараметризации” байесовских моделей. Благодаря относительной простоте, с которой непараметрические распределения могут быть встроены в существующие байесовские модели, а также развитие методов вывода, скорость появления работ, в которых применяются непараметрические методы, непрерывно растет, одновременно с этим расширяется область их применения.

Эволюция методов вывода для непараметрических моделей

[33] В данном случае под противопоставлением следует понимать не соперничество в научной среде приверженцев тех или иных методов, а конструктивное отличие в определении обоих типов моделей и методов работы с ними.

[34] (Sudderth & Jordan, 2008)

[35] (Ghosh, Ungureanu, Sudderth, & Blei, 2011)

[36] (MacKay & Peto, 1995)

[37] (Teh Y. W., 2006)

[38] (Blei & Frazier, 2011)

[39] (Liang, Petrov, Jordan, & Klein, 2007)

[40] (Meeds, Ghahramani, Neal, & Roweis, 2006)

[41] (Paisley & Carin, 2009)

Вполне возможно, что определяющим для развития области оказалось наличие простого и достаточно эффективного метода вывода для процесса китайского ресторана, а именно схемы Гиббса, которую использовал Расмуссен. Несмотря на то, что процесс Дирихле был уже давно известен, для вывода в моделях на его основе применялись малоэффективные и, главное, сложные в реализации и настройке методы генерации выборки с использованием цепей Маркова (англ. Markov Chain Monte Carlo, MCMC) общего типа. Схема Гиббса также является методом из этого семейства, но в отличие от многих других методов может быть очень легко реализована в ряде случаев и не требует настройки при выводе.

По этим причинам схема Гиббса часто используется исследователями в качестве первой процедуры вывода при реализации новой модели. В виду того, что многие байесовские модели сами по себе формулируются, как *генеративный процесс*, то есть, как процесс генерации выборки, распределенной согласно данной модели, схема Гиббса часто естественным образом может быть получена из определения модели. Нередко, как и в случае работы Расмуссена, подробное рассмотрение выкладок, возникающих в схеме Гиббса, наталкивает исследователей на новые идеи[42].

Несмотря на свою простоту схема Гиббса также имеет и ряд недостатков. Как и все методы MCMC, выборка, получаемся по схеме Гиббса, гарантированно сходится к настоящему апостериорному распределению только при бесконечной длине цепи Маркова, причем до сих пор не существует способа определить ее сходимость, поэтому исследователи просто делают столько шагов по схеме, сколько им позволяют их вычислительные возможности. Подобный подход может быть приемлем для теоретической работы и публикации результатов в научных статьях, но не всегда применим в промышленных приложениях, где необходимо быть уверенным в надежности результатов.

Поэтому в настоящий момент активно развиваются так называемые *вариационные методы* байесовского вывода[43], которые приближают апостериорное распределение другим распределением из некоторого достаточно простого семейства, в котором, как правило, распределение на все переменные разделяется на произведение одномерных распределений на каждую из переменных. Построение схем вариационного вывода в общем случае является нетривиальной задачей и зачастую требует применения новых математических идей, из-за чего они существуют не для каждой модели.

[42] (Kulis & Jordan, 2012)
[43] (Lawrence, 2000)

Поскольку в вариационных методах задача вывода формулируется не как задача генерации несмещенной выборки из апостериорного распределения, а как *задача оптимизации* по приближению распределения, то вариационные алгоритмы являются *детерминированными*, то есть, их время работы предсказуемо, и потому они представляют не только теоретический интерес, но и существенно более успешно могут применяться на практике.

В последние годы вариационные методы стали еще более популярны на волне актуальности задач обработки так называемых больших объемов данных (англ. big data). Так, Мэтом Хоффманом, Дэвидом Блеи, одним из ключевых исследователей непараметрических байесовских методов, и другими был предложен метод стохастического вариационного вывода[44], позволяющий осуществлять вывод в достаточно широком классе моделей, не просматривая одновременно всю обучающую выборку. Таким образом, в том числе и для непараметрических моделей вариационный вывод стало возможным осуществлять параллельно и обрабатывать огромные массивы данных в промышленных масштабах. Для непараметрических моделей параллельные процедуры вывода предлагались и ранее[45], но они не содержали в себе единого рецепта, пригодного для достаточно широкого класса моделей. Также подобные разработки появлялись и для схемы Гиббса[4647].

Несмотря на то, что методы МСМС и вариационные методы поначалу неявно противопоставлялись друг-другу, в дальнейшем многие идеи из работ, в которых для вывода использовались именно методы МСМС, были заимствованы для построения новых вариационных алгоритмов. В частности, популярный прием *коллапсирования*, заключающийся в том, что совместное распределение исходной модели интегрируется по части переменных дабы в дальнейшем исключить из процедуры вывода, был впервые задействован в выводе по схеме Гиббса для модели латентного размещения Дирихле[48] и неоднократно доказал свою эффективность в самых разных задачах. Этот прием был столь популярен, что несмотря на определенные математические трудности, он был адаптирован для вариационных аналогов, в том числе и для непараметрических моделей[49].

---

[44] (Hoffman, Blei, Wang, & Paisley, 2013)
[45] (Asuncion, Smyth, & Welling, 2008)
[46] (Williamson , Dubey, & Xing, 2013)
[47] (Gonzalez, Low, Gretton, & Guestrin, 2011)
[48] (Griffiths & Steyvers, 2004)
[49] (Kurihara, Welling, & Teh, 2007)

Кроме того, алгоритм *фильтра частиц*[50], породивший семейство методов последовательной генерации выборки (англ. Sequential Monte Carlo, SMC), будучи одним из методов MCMC, оказал влияние на последние работы по *вариационному выводу в реальном времени* (англ. online inference) для непараметрических моделей на основе процесса Дирихле[51].

Таким образом, очевидно взаимное влияние в некотором смысле противоборствующих направлений в области байесовского вывода. Крайне интересным представляется появление в последнее время *гибридных* методов вывода для непараметрических моделей, с одной стороны представляющих из себя вариационные алгоритмы, а с другой использующих внутри себя методы MCMC. Примером может послужить работа Вонга и Блеи[52], в которой при помощи схемы Гиббса элегантно была решена важная проблема вариационных алгоритмов для процесса Дирихле на основе процесса разлома палки, а именно необходимость вручную ограничивать максимально допустимую размерность модели. Работа Вонга и Блея, а также ряд других работ[53], обратили внимание научного сообщества на проблемы, вызываемые данным ограничением, которыми ранее пренебрегалось.

Заключение

В данном реферате была рассмотрена история развития непараметрических байесовских методов от ранних работ, задолго до появления понятия машинного обучения, до самых последних результатов в области. В настоящий момент непараметрические методы продолжают набирать популярность и активно развиваться, история их применения далеко от своего завершения. В то время как классические методы вроде процесса Дирихле, открытого еще в 1970-ые годы, достаточно хорошо изучены в самых разных аспектах, в отношении более новых процессов подобного понимания еще не достигнуто, не говоря уже о постоянно разрабатываемых новых распределениях. Можно предположить, что в дальнейшем в соответствии с общим трендом развития области машинного обучения будут разрабатываться новые методы, все более специфические и специально заточенные под конкретные задачи, для которых будут требоваться новые, более эффективные методы вывода.

Список литературы

[50] (Gordon, Salmond, & Smith, 1993)
[51] (Lin, 2013)
[52] (Wang & Blei, 2012)
[53] (Bryant & Sudderth, 2012)

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on , 6* (19), 716-723.

Aldous, D. J. (1985). Exchangeability and related topics. *Lecture Notes in Mathematics* (1117), 1-198.

Asuncion, A., Smyth, P., & Welling, M. (2008). Asynchronous Distributed Learning of Topic Models. *Neural Information Processing Systems (NIPS)*.

Beal, M. J., Ghahramani, Z., & Rasmussen, C. E. (2001). The infinite hidden Markov model. *Advances in neural information processing systems* , (стр. 557-584).

Bishop, C. M. (1999). Bayesian PCA. *Advances in neural information processing systems*, (стр. 382-388).

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (T. 1). New York: Springer.

Blei, D. M., & Frazier, P. I. (2011). Distance Dependent Chinese Restaurant Processes. *J. Mach. Learn. Res. , 12*.

Blei, D. M., & Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian analysis , 1* (1), 121-143.

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM) , 2* (57).

Broderick, T., Jordan, M. I., & Pitman, J. (2012). Beta Processes, Stick-Breaking and Power Laws. *Bayesian Analysis , 2* (7), 439-476.

Bryant, M., & Sudderth, E. B. (2012). Truly nonparametric online variational inference for hierarchical Dirichlet processes. *Advances in Neural Information Processing Systems*, (стр. 2708-2716).

de Finetti, B. (1931). Funzione caratteristica di un fenomeno aleatorio. *Atti della R. Academia Nazionale dei Lincei, Serie 6.* (4), 251–299.

Dunson, D. B. (2010). Nonparametric Bayes applications to biostatistics. B L. Hjort, & Nils, *Bayesian nonparametrics.*

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics , 1* (2), 209-230.

Ghosh, S., Ungureanu, A. B., Sudderth, E. B., & Blei, D. M. (2011). Spatial distance dependent Chinese restaurant processes for image segmentation. *Advances in Neural Information Processing Systems*, (стр. 1476-1484).

Gonzalez, J., Low, Y., Gretton, A., & Guestrin, C. (2011). Parallel Gibbs Sampling: From Colored Fields to Thin Junction Trees. *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)* (стр. 324-332). Microtome Publishing.

Gordon, N. J., Salmond, D. J., & Smith, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEEE Proceedings F (Radar and Signal Processing) , 140* (2).

Griffiths, T. L., & Ghahramani, Z. (2005). Infinite latent feature models and the Indian buffet process. *Neural Information Processing Systems.*

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America , 101*, 5228–5235.

Hoffman, M., Blei, D., Wang, C., & Paisley, J. (2013). Stochastic Variational Inference. *Journal of Machine Learning Research* (14), 1303-1347.

Hoppe, F. M. (1984). Pólya-like urns and the Ewens' sampling formula. *Journal of Mathematical Biology , 20* (1), 91-94.

Kulis, B., & Jordan, M. (2012). Revisiting k-means: New Algorithms via Bayesian Nonparametrics. *Proceedings of the 29th International Conference on Machine Learning.* Edinburgh.

Kurihara, K., Welling, M., & Teh, Y. W. (2007). Collapsed Variational Dirichlet Process Mixture Models. *International Joint Conference on Artificial Intelligence*, *7*.

Lawrence, N. D. (2000). *Variational Inference in Probabilistic Models, PhD thesis*. Cambridge: Computer Laboratory, Cambridge University.

Liang, P., Petrov, S., Jordan, M. I., & Klein, D. (2007). The Infinite PCFG Using Hierarchical Dirichlet Processes. *EMNLP-CoNLL*, (стр. 668-697).

Lin, D. (2013). Online Learning of Nonparametric Mixture Models via Sequential Variational Approximation. *Neural Information Processing Systems (NIPS).*

MacEachern, S. (1999). Dependent Nonparametric Processes. *ASA Proceedings of the Section on Bayesian Statistical Science* .

MacKay, D. J., & Peto, L. C. (1995). A hierarchical Dirichlet language model. *Natural language engineering , 3* (1), 289-308.

Meeds, E., Ghahramani, Z., Neal, R. M., & Roweis, S. T. (2006). Modeling Dyadic Data with Binary Latent Factors . *Advances in neural information processing systems*, (стр. 977-984).

Miller, J. W., & Harrison, T. M. (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. *arXiv preprint* .

Paisley, J., & Carin, L. (2009). Nonparametric factor analysis with beta process priors. *Proceedings of the 26th Annual International Conference on Machine Learning* (стр. 777-784). ACM.

Pitman, J., & Yor, M. (1997). The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *Annals of Probability , 2* (25), 855–900.

Pitt, M. K., & Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American statistical association , 94* (466).

Rasmussen, C. E. (2000). The Infinite Gaussian Mixture Model . *Advances in Neural Information Processing Systems.* MIT Press.

Rodriguez, A., Dunson, D. B., & Gelfand, A. E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association , 483* (103).

Ruiz, F., Valera, I., Blanco, C., & Perez-Cruz, F. (2012). Bayesian Nonparametric Modeling of Suicide Attempts. *Advances in Neural Information Processing Systems*, (стр. 1862-1870).

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica , 4*, 639-650.

Sudderth, B. E., & Jordan, M. I. (2008). Shared segmentation of natural scenes using dependent Pitman-Yor processes. *Advances in Neural Information Processing Systems*, (стр. 1585-1592).

Teh, Y. W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (стр. 985-992). Association for Computational Linguistics.

Teh, Y. W., Gorur, D., & Ghahramani, Z. (2007). Stick-breaking construction for the Indian buffet process. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, (стр. 441-449). San Juan, Puerto Rico.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the american statistical association , 101* (476).

Thibaux, R., & Jordan, M. I. (2007). Hierarchical Beta Processes and the Indian Bu®et Process . *International Conference on Artificial Intelligence and Statistics*, (стр. 564-571).

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *The Journal of Machine Learning Research* (1), 211-244.

Vapnik, V. N., & Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications , 2* (16), 264-280.

Wang, C., & Blei, D. (2012). Truncation-free online variational inference for bayesian nonparametric models. *Advances in Neural Information Processing Systems*, (стр. 422-430).

Williamson , S., Dubey, A., & Xing, E. (2013). Parallel Markov Chain Monte Carlo for Nonparametric Mixture Models. *JMLR W&CP , 1* (28).

Xu, Z., Tresp, V., Yu, S., & Yu, K. (2008). Nonparametric relational learning for social network analysis. *KDD'2008 Workshop on Social Network Mining and Analysis*.