От алгебраического подхода к проблеме распознавания Ю.И.Журавлёва к ансамблированию моделей в широком смысле

Воронцов Константин Вячеславович
(ВМК МГУ, Институт ИИ МГУ, МФТИ, ФИЦ ИУ РАН)

IDP-2022

Интеллектуализация Обработки Информации
Москва • 6–9 декабря 2022

Содержание

Ансамблирование предсказательных моделей

$X^\ell = (x_i, y_i)_{i=1}^\ell \subset X \times Y$ — обучающая выборка, $y_i = y^*(x_i)$
$a_t\colon X \to Y,\ \ t = 1, \dots, T$ — обучаемые *базовые алгоритмы*

Идея ансамблирования (Ю.И.Журавлёв): как из множества по отдельности плохих алгоритмов a_t построить один хороший?

Декомпозиция базовых алгоритмов $a_t(x) = C(b_t(x))$
$a_t\colon X \xrightarrow{b_t} R \xrightarrow{C} Y$, где R — удобное *пространство оценок*,
b_t — базовые *алгоритмические операторы*,
C — *решающее правило* простого вида

Ансамбль (композиция) базовых алгоритмов $a_1, \dots, a_T$,
$F\colon R^T \to R$ — *корректирующая* (агрегирующая) *операция*

$$a(x) = C\bigl(F(b_1(x), \dots, b_T(x))\bigr),$$

Ю.И.Журавлёв. Об алгебраическом подходе к решению задач распознавания или классификации. Проблемы кибернетики, 1978.

Обучение предсказательных моделей и их ансамблей

$\mathscr{L}(b, x_i)$ — функция потерь модели $b(x, \omega)$ на объекте x_i
Минимизация эмпирического риска для базовых алгоритмов:

$$\sum_{i=1}^{\ell} \mathscr{L}\bigl(b_t(x_i, \omega), y_i\bigr) \to \min_{\omega}$$

Минимизация эмпирического риска для ансамбля
в пространстве параметров $\Omega = (\omega_1, \ldots, \omega_T, \omega_F)$:

$$\sum_{i=1}^{\ell} \mathscr{L}\Bigl(F\bigl(b_1(x_i, \omega_1), \ldots, b_T(x_i, \omega_T), \omega_F\bigr), y_i\Bigr) \to \min_{\Omega}$$

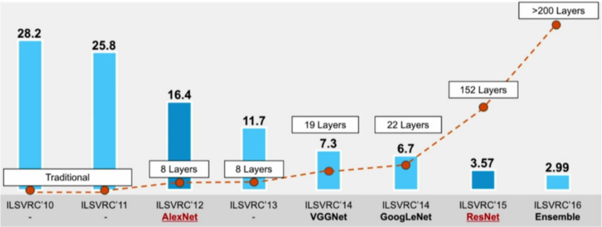
Ю.И.Журавлёв. Корректные алгебры над множествами некорректных (эвристических) алгоритмов (I, II, III). Кибернетика, Киев, 1977–1978.
M.Kearns, L.G.Valiant. Cryptographic limitations on learning Boolean formulae and finite automata. 1989.
Y.Freund, R.E.Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. 1995.
К.В.Рудаков, К.В.Воронцов. О методах оптимизации и монотонной коррекции в алгебраическом подходе к проблеме распознавания. Доклады РАН, 1999.

Философия ансамблирования

Ансамблировать можно только нечто гомогенное.

1. **Декомпозиция** — разделение модели алгоритма a_t на алгоритмический оператор b_t и решающее правило C:
$$a_t = C \circ b_t$$
2. **Гомогенизация** — разнородные модели имеют общее пространство оценок R и общую структуру алгоритмического оператора b_t как отображения
$$b_t \colon X \to R$$
3. **Ансамблирование** — совместное обучение базовых алгоритмических операторов для решения общей задачи:
$$a = C \circ F(b_1, \ldots, b_T)$$

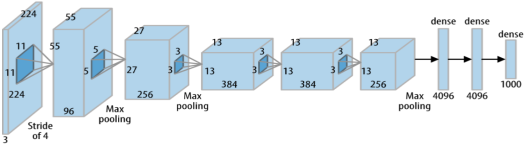
Глубокие свёрточные сети для классификации изображений



Старт в 2009

Человеческий уровень ошибок 5% пройден в 2015

Свёрточные нейронные сети
AlexNet (2012)
ResNet (2015)



Li Fei-Fei et al. ImageNet: A large-scale hierarchical image database. 2009.
Li Fei-Fei et al. Construction and analysis of a large scale image ontology. 2009.
Krizhevsky A., Sutskever I., Hinton G. ImageNet classification with deep convolutional neural networks. 2012.

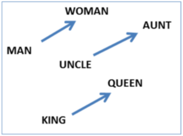
Эволюция подходов машинного обучения в анализе текстов

Декомпозиция задач по уровням пирамиды NLP

- морфологический анализ, лемматизация, опечатки
- синтаксический анализ, выделение терминов, NER
- семантический анализ, выделение фактов, тем

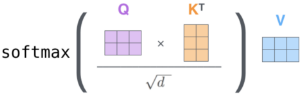
Модели векторных представлений (эмбедингов) слов на основе матричных разложений

- модели дистрибутивной семантики:
  word2vec [Mikolov, 2013], FastText [Bojanowski, 2016]
- тематические модели LDA [Blei, 2003], ARTM [2014]

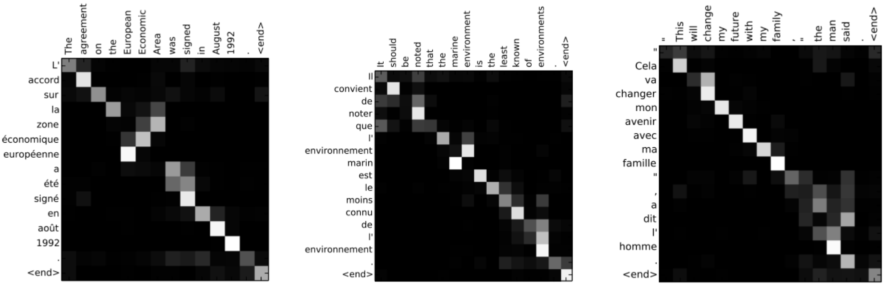


Нейросетевые модели локальных контекстов

- рекуррентные нейронные сети
- модели внимания и трансформеры:
  BERT [2018], GPT-3 [2020] и др.



Модели внимания для машинного перевода



Вход: $\{x_i\}$ — последовательность слов входного языка
Выход: $\{y_t\}$ — последовательность слов выходного языка

Интерпретация: матрица a_{it} показывает, на какие слова x_i модель обращает внимание, генерируя слово перевода y_t

Bahdanau et al. Neural machine translation by jointly learning to align and translate. 2015.
Vaswani et al. Attention is all you need. 2017.
Dichao Hu. An Introductory Survey on Attention Mechanisms in NLP Problems. 2018.

Модели внимания для аннотирования изображений

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

Подсвечены области, на которые модель обращает внимание, когда генерирует подчёркнутое слово в аннотации изображения

Kelvin Xu et al. Show, attend and tell: neural image caption generation with visual attention. 2016

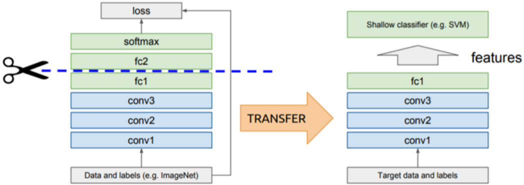
Предобучение (pre-training), перенос обучения (transfer learning)

Обучение модели векторизации $z = f(x, \alpha)$ на выборке $\{x_i\}_{i=1}^{\ell}$:

$$\sum_{i=1}^{\ell} \mathscr{L}_i(g(f(x_i, \alpha), \beta)) \to \min_{\alpha,\beta}$$

Обучение целевой модели $y = g(z, \beta)$ на малых данных:

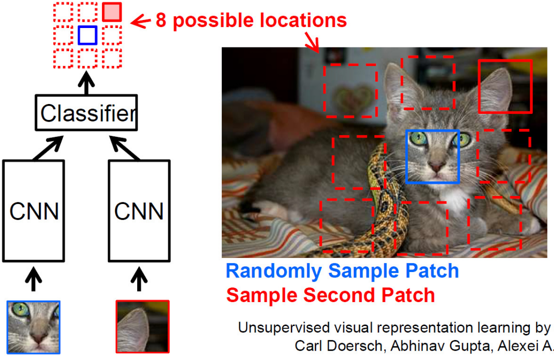
$$\sum_{i=1}^{m} \mathscr{L}'_i(g'(f(x'_i, \alpha), \beta')) \to \min_{\beta'}$$



Sinno Jialin Pan, Qiang Yang. A Survey on Transfer Learning. 2009
J.Yosinski et al. How transferable are features in deep neural networks? 2014.

Самостоятельное обучение (self-supervised learning)

Модель векторизации $z = f(x, \alpha)$ обучается предсказывать взаимное расположение пар фрагментов одного изображения



Unsupervised visual representation learning by context prediction, Carl Doersch, Abhinav Gupta, Alexei A. Efros, ICCV 2015

Преимущество: сеть выучивает векторные представления объектов без размеченной обучающей выборки (без ImageNet).

Многозадачное обучение (multi-task learning)

$z = f(x, \alpha)$ — векторизация, универсальная для всех моделей
$y_t = g_t(z, \beta)$ — специфичная часть модели для задачи $t \in T$

Одновременное обучение модели f по выборкам X^t, $t \in T$:

$$\sum_{t \in T} \sum_{i \in X^t} \mathscr{L}_{ti}\bigl(g_t(f(x_{ti}, \alpha), \beta_t)\bigr) \;\to\; \min_{\alpha, \{\beta_t\}}$$

Обучаемость (learnability): качество решения отдельной задачи $\langle X^t, \mathscr{L}_t, g_t \rangle$ улучшается с ростом объёма выборки $\ell_t = |X^t|$.

Learning to learn: качество решения каждой из задач $t \in T$ улучшается с ростом как ℓ_t, так и общего числа задач $|T|$.

Few-shot learning: для решения новой задачи t достаточно небольшого числа примеров, иногда даже одного.

M.Crawshaw. Multi-task learning with deep neural networks: a survey. 2020
Y.Wang et al. Generalizing from a few examples: a survey on few-shot learning. 2020

Философия многозадачного обучения

1. **Декомпозиция** — разделение моделей $y_t\colon X \to Y_t$ на векторизатор $z = f(x, \alpha)$ и предиктор $y_t = g_t(z, \beta)$:

$$y_t(x) = g_t(f(x, \alpha), \beta_t)$$

2. **Гомогенизация** — разнородные модели имеют общий векторизатор $z = f(x, \alpha)$ и общее векторное пространство представлений (эмбедингов) Z:

$$f\colon X \to Z$$

3. **Ансамблирование** — совместное обучение общего векторизатора для решения разнородных задач:

$$\sum_{t \in T} \sum_{i \in X^t} \mathscr{L}_{ti}\bigl(g_t(f(x_{ti}, \alpha), \beta_t)\bigr) \ \to \ \min_{\alpha, \{\beta_t\}}$$

Автокодировщики — векторизаторы, обучаемые без учителя

Дано: обучающая выборка объектов $\{x_i\}_{i=1}^{\ell}$

Найти: $z = f(x, \alpha)$ — модель кодировщика (encoder)

$\hat{x} = g(z, \beta)$ — модель декодировщика (decoder)

Критерий: качество реконструкции исходных объектов

$$\sum_{i=1}^{\ell} \mathscr{L}\big(g(f(x_i, \alpha), \beta), x_i\big) \to \min_{\alpha,\beta}$$

Квадратичная функция потерь: $\mathscr{L}(\hat{x}, x) = \|\hat{x} - x\|^2$

Пример 1. Линейный автокодировщик: $x \in \mathbb{R}^n$, $z \in \mathbb{R}^m$

$$f(x, A) = \underset{m\times n}{A}\, x, \qquad g(z, B) = \underset{n\times m}{B}\, z$$

Пример 2. Двухслойная сеть с функциями активации σ_f, σ_g:

$$f(x, A) = \sigma_f(Ax + a), \qquad g(z, B) = \sigma_g(Bz + b)$$

Автокодировщики для векторизации и обучения с учителем

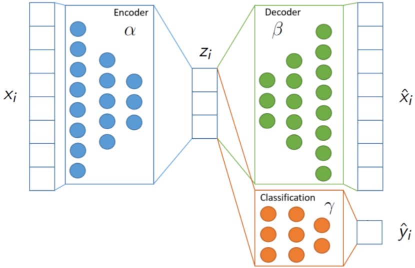
Данные: размеченные $(x_i, y_i)_{i=1}^{k}$, неразмеченные $(x_i)_{i=k+1}^{\ell}$

Найти:

$z_i = f(x_i, \alpha)$ — кодировщик

$\hat{x}_i = g(z_i, \beta)$ — декодировщик

$\hat{y}_i = \hat{y}(z_i, \gamma)$ — предиктор



Функции потерь:

$\mathscr{L}(\hat{x}_i, x_i)$ — реконструкция

$\tilde{\mathscr{L}}(\hat{y}_i, y_i)$ — предсказание

Критерий: совместное обучение автокодировщика и предсказательной модели (классификации, регрессии или др.):

$$\sum_{i=1}^{\ell} \mathscr{L}\big(g(f(x_i, \alpha), \beta), x_i\big) + \lambda \sum_{i=1}^{k} \tilde{\mathscr{L}}\big(\hat{y}(f(x_i, \alpha), \gamma), y_i\big) \to \min_{\alpha, \beta, \gamma}$$

Dor Bank, Noam Koenigstein, Raja Giryes. Autoencoders. 2020

Задача тематического моделирования

Дано: коллекция текстовых документов

- W — конечное множество термов (слов, токенов)
- D — конечное множество документов
- n_{dw} — частота терма w в документе d

Найти: вероятностную тематическую модель

$$p(w|d) = \sum_{t \in T} p(w \,|\, \cancel{d,}\, t)\, p(t|d) = \sum_{t \in T} \phi_{wt} \theta_{td}$$

где $\phi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$ — параметры модели

Критерий: максимум логарифма правдоподобия

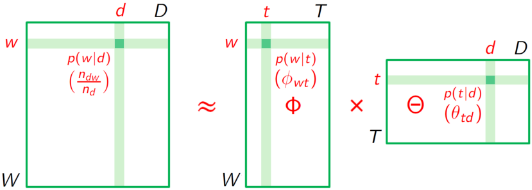
$$\ln \prod_{d,w} p(w|d)^{n_{dw}} = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi,\Theta}$$

при ограничениях $\phi_{wt} \geqslant 0$, $\sum_w \phi_{wt} = 1$, $\theta_{td} \geqslant 0$, $\sum_t \theta_{td} = 1$

Hofmann T. Probabilistic Latent Semantic Indexing. ACM SIGIR, 1999.

Три интерпретации задачи тематического моделирования

1. **Мягкая кластеризация** документов по кластерам-темам
2. Низкоранговое стохастическое **матричное разложение:**



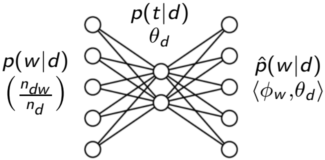
3. **Автокодировщик** документов в тематические эмбединги:

кодировщик $\quad f_\Phi\colon \frac{n_{dw}}{n_d} \to \theta_d$

декодировщик $\quad g_\Phi\colon \theta_d \to \Phi\theta_d$

задача реконструкции:

$$\sum_d \mathrm{KL}\left(\frac{n_{dw}}{n_d} \,\Big\|\, \langle \phi_w, \theta_d \rangle\right) \to \min_{\Phi,\Theta}$$



Байесовское обучение — основной подход в Topic Modeling

$X = (d_i, w_i)_{i=1}^n$ — наблюдаемые переменные, коллекция длины n
$Z = (t_i)_{i=1}^n$ — скрытые переменные
$\Omega = (\Phi, \Theta)$ — искомые параметры модели
$\gamma = (\beta, \alpha)$ — гиперпараметры априорных распределений

Задача байесовского вывода — получить не Ω, а $p(\Omega|X, \gamma)$

Вариационный байесовский вывод:

вывести $p(Z, \Omega|X, \gamma) \propto p(X, Z|\Omega, \gamma)\, p(\Omega|\gamma)$

Сэмплирование Гиббса:

вывести $p(Z|X, \gamma)$
сэмплировать $Z \sim p(Z|X, \gamma)$
вывести $p(\Omega|X, Z, \gamma) \propto p(X, Z|\Omega, \gamma)\, p(\Omega|\gamma)$

Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. JMLR, 2003.

Общий взгляд на байесовское обучение, MAP и ARTM

Байесовский вывод апостериорного распределения $p(\Omega|X)$ (сложный, приближённый) ради получения точечной оценки Ω:

$$\mathrm{Posterior}(\Omega|X,\gamma) \propto p(X|\Omega)\,\mathrm{Prior}(\Omega|\gamma)$$

$$\Omega := \arg\max_{\Omega} \mathrm{Posterior}(\Omega|X,\gamma)$$

Максимизация апостериорной вероятности (MAP) даёт точечную оценку Ω напрямую, без вывода Posterior:

$$\Omega := \arg\max_{\Omega} \bigl(\ln p(X|\Omega) + \underbrace{\ln \mathrm{Prior}(\Omega|\gamma)}_{R(\Omega)}\bigr)$$

Многокритериальная аддитивная регуляризация (ARTM) обобщает MAP на любые регуляризаторы и их комбинации:

$$\Omega := \arg\max_{\Omega} \Bigl(\ln p(X|\Omega) + \sum_k \lambda_k R_k(\Omega)\Bigr)$$

ARTM: аддитивная регуляризация тематических моделей

Максимизация log правдоподобия с регуляризатором R:

$$\sum_{d,w} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}; \qquad R(\Phi,\Theta) = \sum_k \tau_k R_k(\Phi,\Theta)$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} \equiv p(t|d,w) = \mathop{\mathrm{norm}}_{t\in T}\bigl(\phi_{wt}\theta_{td}\bigr) \\ \text{M-шаг:} & \phi_{wt} = \mathop{\mathrm{norm}}_{w\in W}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right), \quad n_{wt} = \sum_{d\in D} n_{dw}p_{tdw} \\ & \theta_{td} = \mathop{\mathrm{norm}}_{t\in T}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right), \quad n_{td} = \sum_{w\in d} n_{dw}p_{tdw} \end{cases}$$

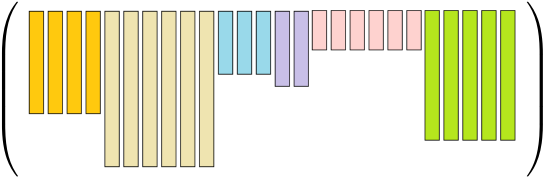
где $\mathop{\mathrm{norm}}_{t\in T}(x_t) = \frac{\max\{x_t,0\}}{\sum_{s\in T}\max\{x_s,0\}}$ — операция нормирования вектора.

Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН, 2014.

Задача максимизации функции на единичных симплексах

Пусть $\Omega = (\omega_j)_{j \in J}$ — набор нормированных неотрицательных векторов $\omega_j = (\omega_{ij})_{i \in I_j}$ различных размерностей $|I_j|$:

$\Omega =$



Задача максимизации функции $f(\Omega)$ на единичных симплексах:

$$\begin{cases} f(\Omega) \to \max\limits_{\Omega}; \\ \sum\limits_{i \in I_j} \omega_{ij} = 1, \quad j \in J; \\ \omega_{ij} \geqslant 0, \quad i \in I_j, \quad j \in J. \end{cases}$$

Теорема о максимизации на единичных симплексах

Операция нормировки вектора: $p_i = \underset{i \in I}{\mathrm{norm}}(x_i) = \dfrac{\max(x_i, 0)}{\sum_k \max(x_k, 0)}$

Теорема (необходимые условия экстремума)

Пусть $f(\Omega)$ непрерывно дифференцируема по Ω.
Если ω_j — вектор локального экстремума задачи $f(\Omega) \to \max$ и $\exists i$: $\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} > 0$, то ω_j удовлетворяет системе уравнений

$$\omega_{ij} = \underset{i \in I_j}{\mathrm{norm}} \left(\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} \right).$$

- Численное решение системы — методом простых итераций
- Итерации похожи на градиентную оптимизацию, но учитывают ограничения и не требуют подбора шага η:

$$\omega_{ij} := \omega_{ij} + \eta \frac{\partial f}{\partial \omega_{ij}}$$

Доказательство теоремы о максимизации на симплексах

Задача: $f(\Omega) \to \max\limits_{\Omega}; \quad \sum\limits_{i \in I_j} \omega_{ij} = 1, \quad \omega_{ij} \geqslant 0, \quad i \in I_j, \quad j \in J.$

Функция Лагранжа:

$$\mathscr{L}(\Omega; \mu, \lambda) = f(\Omega) + \sum_{j \in J} \lambda_j \Big(\sum_{i \in I_j} \omega_{ij} - 1 \Big) - \sum_{j \in J} \sum_{i \in I_j} \mu_{ij} \omega_{ij}.$$

Условия Каруша–Куна–Таккера для вектора ω_j:

$$\frac{\partial f(\Omega)}{\partial \omega_{ij}} = \lambda_j - \mu_{ij}; \quad \mu_{ij}\omega_{ij} = 0.$$

Умножим обе части первого равенства на ω_{ij}:

$$A_{ij} \equiv \omega_{ij} \frac{\partial f(\Omega)}{\partial \omega_{ij}} = \omega_{ij}\lambda_j.$$

Согласно условию теоремы $\exists i$: $A_{ij} > 0$. Значит, $\lambda_j > 0$.

Если $\frac{\partial f(\Omega)}{\partial \omega_{ij}} < 0$ для некоторого i, то $\mu_{ij} > 0 \;\Rightarrow\; \omega_{ij} = 0$.

Тогда $\omega_{ij}\lambda_j = (A_{ij})_+$; $\lambda_j = \sum\limits_i (A_{ij})_+ \;\Rightarrow\; \omega_{ij} = \operatorname*{norm}\limits_i (A_{ij})$.

■

Теорема о сходимости итерационного процесса

$$\omega_{ij}^{t+1} = \operatorname*{norm}_{i \in I_j} \left(\omega_{ij}^t \frac{\partial f(\Omega^t)}{\partial \omega_{ij}^t} \right)$$

Теорема. Пусть $f(\Omega)$ — ограниченная сверху, непрерывно дифференцируемая функция, и все Ω^t, начиная с некоторой итерации t^0 обладают свойствами:

- $\forall j \in J \;\; \forall i \in I_j \;\; \omega_{ij}^t = 0 \to \omega_{ij}^{t+1} = 0$ (сохранение нулей)
- $\exists \varepsilon > 0 \;\; \forall j \in J \;\; \forall i \in I_j \;\; \omega_{ij}^t \notin (0, \varepsilon)$ (отделимость от нуля)
- $\exists \delta > 0 \;\; \forall j \in J \;\; \exists i \in I_j \;\; \omega_{ij}^t \frac{\partial f(\Omega^t)}{\partial \omega_{ij}} \geqslant \delta$ (невырожденность)

Тогда $f(\Omega^{t+1}) > f(\Omega^t)$ и $\left|\omega_{ij}^{t+1} - \omega_{ij}^t\right| \to 0$ при $t \to \infty$.

Ирхин И. А., Воронцов К. В. Сходимость алгоритма аддитивной регуляризации тематических моделей // Труды Института математики и механики УрО РАН. 2020.

Дебайесизация и комбинирование тематических моделей

«Представляется важной задача освобождения всюду, где это возможно, от излишних вероятностных допущений»
(А. Н. Колмогоров, 1987)

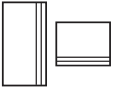
В результате *дебайесизации* в тематическом моделировании

- **теряется:**
  — возможность оценивания апостериорных распределений (которая практически никогда и не использовалась)
- **приобретается:**
  — более общее доказательство сходимости
  — более удобная формализация широкого класса моделей
  — возможность комбинирования моделей (ARTM)
  — возможность модульной реализации (BigARTM)

Переход к байесовской регуляризации минуя классическую надолго заблокировал развитие комбинированных ВТМ

Регуляризаторы для улучшения интерпретируемости тем

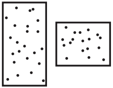
background



Сглаживание фоновых тем $B \subset T$:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in B} \sum_{w} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d} \sum_{t \in B} \alpha_t \ln \theta_{td}$$

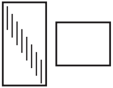
sparse



Разреживание предметных тем $S = T \setminus B$:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in S} \sum_{w} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d} \sum_{t \in S} \alpha_t \ln \theta_{td}$$

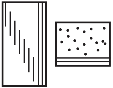
decorrelated



Декоррелирование для повышения различности тем:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s} \sum_{w} \phi_{wt} \phi_{ws}$$

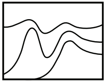
interpretable



Сглаживание + разреживание + декоррелирование для улучшения интерпретируемости тем

Регуляризаторы для учёта дополнительной информации

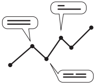
temporal



Темпоральные модели с модальностью времени i:

$$R(\Phi) = -\tau \sum_{i \in I} \sum_{t \in T} \left| \phi_{it} - \phi_{i-1,t} \right|$$

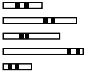
regression



Линейная модель регрессии $\hat{y}_d = \langle v, \theta_d \rangle$ документов:

$$R(\Theta, v) = -\tau \sum_{d \in D} \Big(y_d - \sum_{t \in T} v_t \theta_{td} \Big)^2$$

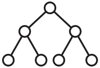
coherence



Модели сочетаемости слов (n_{uv} — частота битерма):

$$R(\Phi) = \tau \sum_{u \in W} \sum_{v \in W} n_{uv} \ln \sum_{t \in T} n_t \phi_{ut} \phi_{vt}$$

hierarchy

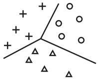


Связь родительских тем t с дочерними подтемами s:

$$R(\Phi, \Psi) = \tau \sum_{t \in T} \sum_{w \in W} n_{wt} \ln \sum_{s \in S} \phi_{ws} \psi_{st}$$

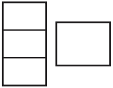
Регуляризаторы для мультимодальных тематических моделей

supervised



Модальности меток классов, категорий или тегов для классификации/категоризации/тегирования текстов

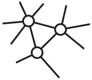
multilanguage



Модальность языков и регуляризация со словарём $\pi_{uwt} = p(u|w,t)$ переводов с языка k на ℓ:

$$R(\Phi,\Pi) = \tau \sum_{u\in W^k} \sum_{t\in T} n_{ut} \ln \sum_{w\in W^\ell} \pi_{uwt}\phi_{wt}$$

graph



Модальность вершин графа v, содержащих D_v:

$$R(\Phi) = -\frac{\tau}{2} \sum_{(u,v)\in E} S_{uv} \sum_{t\in T} n_t^2 \left(\frac{\phi_{vt}}{|D_v|} - \frac{\phi_{ut}}{|D_u|}\right)^2.$$

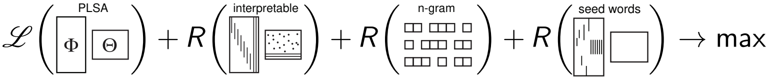
geospatial

Модальность геолокаций g с близостью $S_{gg'}$:

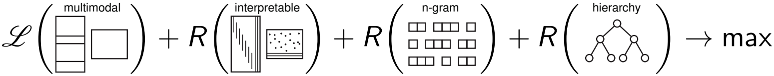
$$R(\Phi) = -\frac{\tau}{2} \sum_{g,g'\in G} S_{gg'} \sum_{t\in T} n_t^2 \left(\frac{\phi_{gt}}{n_g} - \frac{\phi_{g't}}{n_{g'}}\right)^2$$

Ансамблирование регуляризаторов в прикладных задачах

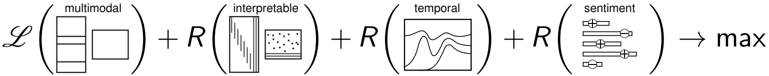
Выявления этнорелевантного дискурса в социальных сетях:



Тематический поиск научных и научно-популярных статей:



Прослеживание событий и мнений в новостных потоках:



M.Apishev et al. Mining ethnic content online with additively regularized topic models, 2016.
A.Ianina, K.Vorontsov. Hierarchical interpretable topical embeddings for exploratory search and real-time document tracking, 2020.
D.Feldman, T.Sadekova, K.Vorontsov. Combining facts, semantic roles and sentiment lexicon in a generative model for opinion mining. 2020

Философия аддитивной регуляризации (ARTM)

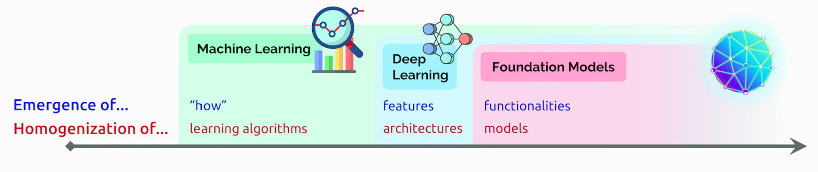
1. **Декомпозиция** — разделение критерия обучения модели на основной (log-правдоподобие) и регуляризатор R:
$$\ln p(X|\Phi,\Theta) + R(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}$$

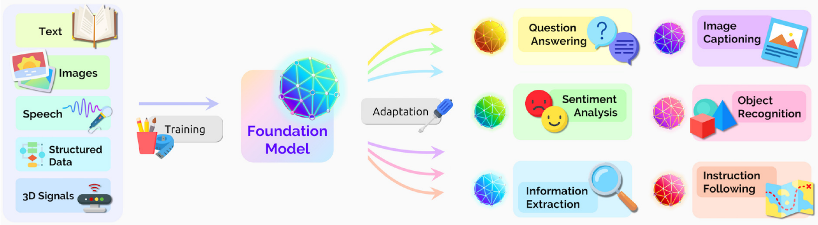
2. **Гомогенизация** — разнородные модели имеют общую структуру векторизатора, матричного разложения и общий основной критерий (log-правдоподобие):
$$f_\Phi\colon X \to \Theta, \qquad \ln p(X|\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}$$

3. **Ансамблирование** — совместное использование регуляризаторов R_k, взятых от разнородных моделей:
$$\ln p(X|\Phi,\Theta) + \sum_k \lambda_k R_k(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}$$

Гомогенизация векторных моделей (Foundation Models)

Обучаемая векторизация данных — глобальный тренд AI/ML





R.Bommasani et al. (Center for Research on Foundation Models, Stanford University)
On the opportunities and risks of foundation models // CoRR, 20 August 2021.

Философия фундаментальных моделей

1. **Декомпозиция** — разделение моделей $y_t\colon X_t \to Y_t$ на векторизатор $z = f(x_t, \alpha_t)$ и предиктор $y_t = g(z_t, \beta_t)$:

$$y_t(x) = g_t(f(x_t, \alpha_t), \beta_t)$$

2. **Гомогенизация** — разнородные модели в разнородных задачах имеют общее пространство эмбедингов Z:

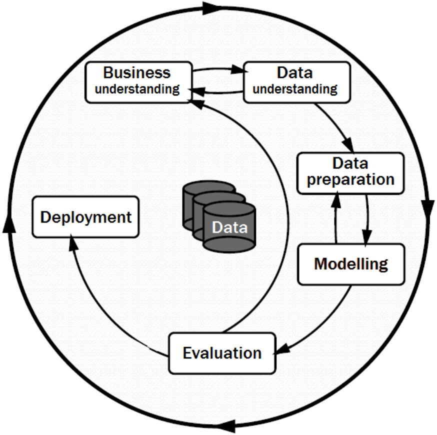
$$f_t\colon X_t \to Z$$

3. **Ансамблирование** — совместное обучение эмбедингов в едином семантическом пространстве для решения разнородных задач:

$$\sum_{t\in T}\sum_{i\in X^t} \mathscr{L}_{ti}\bigl(g_t(f(x_{ti}, \alpha_t), \beta_t)\bigr) \to \min_{\{\alpha_t, \beta_t\}}$$

Межотраслевой стандарт интеллектуального анализа данных

CRISP-DM: CRoss Industry Standard Process for Data Mining (1999)



Компании-инициаторы:

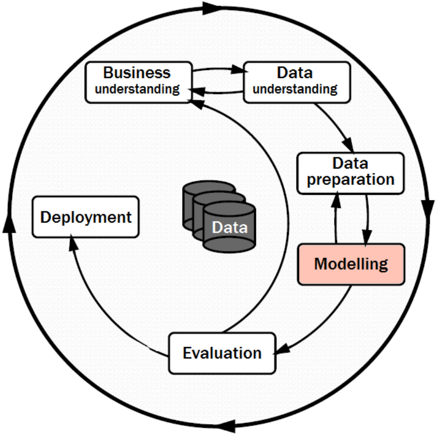
- SPSS
- Teradata
- Daimler AG
- NCR Corp.
- OHRA

Шаги процесса:

- понимание бизнеса
- понимание данных
- предобработка данных и инженерия признаков
- разработка моделей и настройка параметров
- оценивание качества
- внедрение

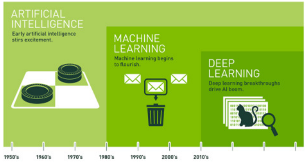
Автоматизация шагов CRISP-DM и эволюция ИИ

CRISP-DM: CRoss Industry Standard Process for Data Mining (1999)



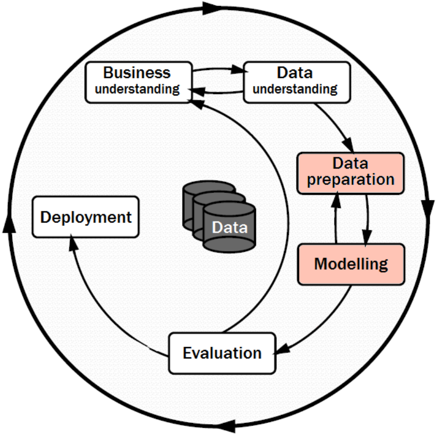
Эволюция ИИ:

- *Expert Systems:* жёсткие модели, основанные на правилах
- *Machine Learning:* параметрические модели, обучаемые по данным



Автоматизация шагов CRISP-DM и эволюция ИИ

CRISP-DM: CRoss Industry Standard Process for Data Mining (1999)

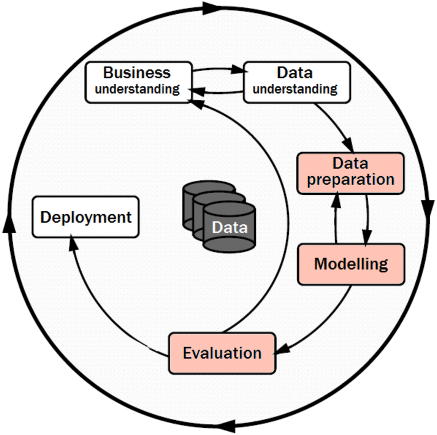


Эволюция ИИ:

- *Expert Systems:* жёсткие модели, основанные на правилах
- *Machine Learning:* параметрические модели, обучаемые по данным
- *Deep Learning:* модели с обучаемой векторизацией данных

Автоматизация шагов CRISP-DM и эволюция ИИ

CRISP-DM: CRoss Industry Standard Process for Data Mining (1999)

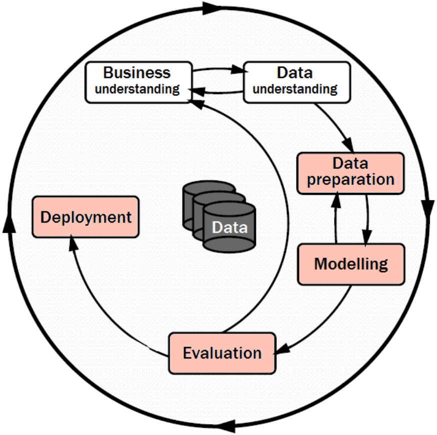


Эволюция ИИ:

- *Expert Systems:* жёсткие модели, основанные на правилах
- *Machine Learning:* параметрические модели, обучаемые по данным
- *Deep Learning:* модели с обучаемой векторизацией данных
- *AutoML:* автоматический выбор моделей и архитектур

Автоматизация шагов CRISP-DM и эволюция ИИ

CRISP-DM: CRoss Industry Standard Process for Data Mining (1999)



Эволюция ИИ:

- *Expert Systems:* жёсткие модели, основанные на правилах
- *Machine Learning:* параметрические модели, обучаемые по данным
- *Deep Learning:* модели с обучаемой векторизацией данных
- *AutoML:* автоматический выбор моделей и архитектур
- *Lifelong Learning:* бесшовная интеграция обучения и выбора моделей в бизнес-процесс

Резюме

- Ансамблирование в широком смысле —
  это *совместное использование* моделей
- Ансамблировать можно только нечто гомогенное
- Общая философия
  — алгебраического подхода Ю.И.Журавлёва,
  — многозадачного обучения,
  — аддитивной регуляризации,
  — фундаментальных моделей:
  «декомпозиция → гомогенизация → ансамблирование»
- Декомпозиция ВТМ потребовала дебайесизации моделей
- ВТМ — это скорее векторизация графов, чем раздел анализа текстов или байесовского обучения, как принято считать
- Теперь это «теория одной теоремы»

К.В.Воронцов. Обзор вероятностных тематических моделей. 2022. – NEW!
http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf

Интерпретируемость тематических векторов

Тематические векторные представления (эмбединги) текста:

- $p(t|d) = \theta_{td}$ для документа d
- $p(t|w) = \phi_{wt}\frac{p(t)}{p(w)}$ для терма w
- $p(t|d, w)$ для локального контекста (d, w)
- $p(t|x)$ для нетекстового объекта x

Интерпретируемость тематических векторов:

- каждая тема t описывается *семантическим ядром* — частотным словарём слов $\bigl\{w\colon p(w|t) > \gamma p(w)\bigr\}$, встречающихся в данной теме в γ раз чаще обычного
- любой объект x с вектором $p(t|x)$ описывается частотным словарём слов $\Bigl\{w\colon p(w|x) = \sum\limits_{t\in T} p(w|t)p(t|x) > \gamma p(w)\Bigr\}$

Цели и не-цели тематического моделирования

Цели:

- Выяснить тематическую кластерную структуру текстовой коллекции, сколько в ней тем и какие они
- Получать интерпретируемые тематические векторные представления (эмбединги) документов, фрагментов, слов $p(t|d)$, $p(t|w)$, $p(t|d,w)$ и нетекстовых объектов $p(t|x)$
- Решать задачи поиска, категоризации, сегментации, суммаризации с помощью тематических эмбедингов

Не-цели:

- Угадывать следующие слова (ТМ — слабые модели языка)
- Генерировать связный текст
- Понимать смысл текста

Модульный подход ARTM: сравнение с байесовским подходом

Для построения композитных моделей в ARTM не нужны ни математические выкладки, ни программирование «с нуля».

<table>
<tr><th>Этапы моделирования</th><th>Bayesian TM</th><th colspan="2">ARTM</th></tr>
<tr><td></td><td>Анализ требований</td><td colspan="2">Анализ требований</td></tr>
<tr><td>Формализация:</td><td>Вероятностная модель порождения данных</td><td>Стандартные критерии</td><td>Свои критерии</td></tr>
<tr><td>Алгоритмизация:</td><td>Байесовский вывод для данной порождающей модели (VI, GS, EP)</td><td colspan="2">Единый регуляризованный EM-алгоритм для любых моделей и их композиций</td></tr>
<tr><td>Реализация:</td><td>Исследовательский код (Matlab, Python, R)</td><td colspan="2">Промышленный код BigARTM (C++, Python API)</td></tr>
<tr><td>Оценивание:</td><td>Исследовательские метрики, исследовательский код</td><td>Стандартные метрики</td><td>Свои метрики</td></tr>
<tr><td></td><td>Внедрение</td><td colspan="2">Внедрение</td></tr>
</table>

-- *нестандартизуемые этапы, уникальная разработка для каждой задачи*

-- *стандартизуемые этапы*

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Большие данные: коллекция не хранится в памяти
- Самый быстрый онлайновый параллельный ARTM
- Встроенная библиотека регуляризаторов и мер качества

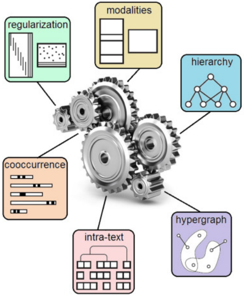
Сообщество:

- Открытый код https://github.com/bigartm
  (discussion group, issue tracker, pull requests)
- Документация http://bigartm.org

Лицензия и среда разработки:

- Свободная коммерческая лицензия (BSD 3-Clause)
- Кросс-платформенность: Linux, MacOS, Windows (32/64 bit)
- Интерфейсы API: C++, Python, командная строка

Ключевые возможности библиотек BigARTM и TopicNet



BigARTM (с 2014 г.)

- библиотека регуляризаторов
- мультимодальные модели
- иерархические модели
- гиперграфовые модели
- модели связности текста

TopicNet (с 2020 г.)

- Перебор сценариев регуляризации для выбора моделей
- Автоматическое протоколирование экспериментов
- Построение «банка тем» из множества моделей
- Визуализация тематических моделей

V.Bulatov, E.Egorov, E.Veselova, D.Polyudova, V.Alekseev, A.Goncharov, K.Vorontsov. TopicNet: making additive regularization for topic modelling accessible. LREC-2020