МАТЕМАТИЧЕСКИЕ ОСНОВЫ ТЕОРИИ ПРОГНОЗИРОВАНИЯ

Лектор

Сенько Олег Валентинович

Лекция 4

Байесовское обучение

Ранее было показано, что максимальную точность распознавания обеспечивает байесовское решающее правило, относящее распознаваемый объект, описываемый вектором переменных (признаков) $X_1, \ldots, X_n$ к классу K_*, для которого условная вероятность $\mathbf{P}(K_* \mid \mathbf{x})$ максимальна.

$\mathbf{x}$

Байесовские методы обучения основаны на аппроксимации условных вероятностей классов в точках признакового пространства с использованием формулы Байеса.

Байесовское обучение

Рассмотрим задачу распознавания классов $K_1, \ldots, K_L$

Формула Байеса позволяет рассчитать условные вероятности классов в точке признакового пространства:

$$\mathbf{P}(K_i \mid \mathbf{x}) = \frac{f_i(\mathbf{x})\mathbf{P}(K_i)}{\sum_{j=1}^{L} f_i(\mathbf{x})\mathbf{P}(K_i)}$$

, где $f_i(\mathbf{x})$ - плотность распределения вероятности для класса K_i;

$\mathbf{P}(K_i)$ - вероятность класса K_i безотносительно к признаковым описаниям (априорная вероятность).

Байесовское обучение

При этом в качестве оценок априорных вероятностей $\mathbf{P}(K_1),\ldots,\mathbf{P}(K_L)$ могут быть взяты доли объектов соответствующих классов в обучающей выборке.

Плотности вероятностей $f_1(\mathbf{x}),\ldots, f_L(\mathbf{x})$ восстанавливаются исходя из предположения об их принадлежности фиксированному типу распределения.

Чаще всего используется многомерное нормальное распределения.

Байесовское обучение

Плотность данного распределения в общем виде представляется выражением

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\,|\Sigma|^{1/2}} \exp[-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})^{t}] \quad ,$$

где $\boldsymbol{\mu}$ - математическое ожидание вектора признаков $\mathbf{x}$;

$\boldsymbol{\Sigma}$ - матрица ковариаций признаков $X_1, \ldots, X_n$;

$|\boldsymbol{\Sigma}|$ - детерминант матрицы $\boldsymbol{\Sigma}$.

Байесовское обучение

Для построения распознающего алгоритма достаточно оценить вектора математических ожиданий $\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_L$ и матрицы ковариаций $\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_L$ для классов $K_1,\ldots,K_L$ соответственно.

Оценка $\boldsymbol{\mu}_i$ вычисляется как среднее значение векторов признаков по объектам обучающей выборки из класса K_i:

$\hat{\boldsymbol{\mu}}_i = \frac{1}{m_i} \sum_{s_j \in \tilde{S}_t \cap K_i} \mathbf{x}_j$, где m_i - число объектов класса K_i в обучающей выборке.

Байесовское обучение

Элемент матрицы ковариаций для класса K_i вычисляется по формуле

$$\sigma^i_{kk'} = \frac{1}{m_i} \sum_{s_j \in \tilde{S}_t \cap K_i} (x_{jk} - \mu^i_k)(x_{jk'} - \mu^i_{k'}), \quad k, k' \in \{1, \ldots, n\}$$

, где μ^i_k - k-ая компонента вектора $\boldsymbol{\mu}^i$. Матрицу, состоящую из элементов $\sigma^i_{kk'}$

Очевидно, что согласно формуле Байеса максимум $\mathbf{P}(K_i \mid \mathbf{x})$ достигается для тех же самых классов для которых максимально произведение $f_i(\mathbf{x})\mathbf{P}(K_i)$.

Байесовское обучение

Очевидно, что для байесовской классификации может использоваться также натуральный логарифм $\ln[f_i(\mathbf{x})\mathbf{P}(K_i)]$ который согласно вышеизложенному может быть оценён выражением $g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}\hat{\mathbf{\Sigma}}_i\mathbf{x}^t) + \mathbf{w}_i\mathbf{x}^t + g_i^0$, , где $\mathbf{w}_i = \hat{\mathbf{\Sigma}}_i^{-1}\hat{\boldsymbol{\mu}}_i$

$g_i^0 = -\frac{1}{2}(\hat{\boldsymbol{\mu}}_i\hat{\mathbf{\Sigma}}_i^{-1}\hat{\boldsymbol{\mu}}_i^t) - \frac{1}{2}\ln(|\hat{\mathbf{\Sigma}}_i|) + \ln(\nu_i) - \frac{n}{2}\ln(2\pi)$ - не зависящее от $\mathbf{x}$ слагаемое;

ν_i - доля объектов класса K_i в обучающей выборке.

Байесовское обучение

Таким образом объект с признаковым описанием $\mathbf{x}$ будет отнесён построенной выше аппроксимацией байесовского классификатора к классу, для которого оценка $g_i(\mathbf{x})$ является максимальной. Следует отметить, что построенный классификатор в общем случае является квадратичным по признакам. Однако классификатор превращается в линейный, если оценки ковариационных матриц разных классов оказываются равными.

Линейный дискриминант Фишера

Рассмотрим вариант метода Линейный дискриминант Фишера (ЛДФ) для распознавания двух классов K_1 и K_2.

В основе метода лежит поиск в многомерном признаковом пространстве такого направления $\mathbf{w}$, чтобы средние значения проекции на него объектов обучающей выборки из классов K_1 и K_2 максимально различались.

Проекцией произвольного вектора $\mathbf{x}$ на направление $\mathbf{w}$ является отношение $\frac{(\mathbf{w}\mathbf{x}^t)}{|\mathbf{w}|}$.

Линейный дискриминант Фишера

В качестве меры различий проекций классов на $\mathbf{w}$ используется функционал

- $$\Phi(\mathbf{w}) = \frac{[\hat{\mathbf{X}}_{w1}(\mathbf{w}) - \hat{\mathbf{X}}_{w2}(\mathbf{w})|}{\hat{d}_1(\mathbf{w}) + \hat{d}_2(\mathbf{w})}$$, где

- $\hat{X}_{wi}(\mathbf{w}) = \frac{1}{m_i} \sum_{s_j \in \tilde{S}_t \cap K_i} \frac{(\mathbf{w}\mathbf{x}_j^t)}{|\mathbf{w}|}$ - среднее значение проекции

векторов, описывающих объекты из класса $K_i, \quad i \in \{1,2\}$

Линейный дискриминант Фишера

- $\hat{d}_i(\mathbf{w}) = \frac{1}{m_i} \sum_{s_j \in \tilde{S}_t \cap K_i} [\frac{(\mathbf{w}\mathbf{x}_j^t)}{|\mathbf{w}|} - \hat{X}_{wi}]^2$ - выборочная

дисперсия проекций векторов, описывающих объекты из класса.

Смысл функционала $\Phi(\mathbf{w})$ ясен из его структуры. Он является по сути квадратом отличия между средними значениями проекций классов на направление $\mathbf{w}$, нормированным на сумму внутриклассовых выборочных дисперсий.

Линейный дискриминант Фишера

Можно показать, что $\Phi(\mathbf{w})$ достигает максимума $\mathbf{w}^t = \hat{\Sigma}_{12}^{-1}(\hat{\boldsymbol{\mu}}_1^t - \hat{\boldsymbol{\mu}}_2^t)$, где $\hat{\Sigma}_{12} = \hat{\Sigma}_1 + \hat{\Sigma}_2$. Таким образом оценка направления, оптимального для распознавания K_1 и K_2 может быть записана в виде $\hat{\mathbf{w}}^t = \hat{\Sigma}_{12}^{-1}(\hat{\boldsymbol{\mu}}_1^t - \hat{\boldsymbol{\mu}}_2^t)$

Распознавание нового объекта s_* по признаковому $\mathbf{x}_*$ описанию производится по величине проекции $\gamma(\mathbf{x}_*) = \dfrac{(\mathbf{w}\mathbf{x}_*^t)}{|\mathbf{w}|}$ с помощью простого порогового правила

При $\gamma(\mathbf{x}_*) \geq b$ объект s_* относится к классу K_1 и s_* относится к классу K_2 в противном случае.

Линейный дискриминант Фишера

Граничный параметр b подбирается по обучающей выборке таким образом, чтобы проекции объектов разных классов на оптимальное направление $\mathbf{w}$ оказались бы максимально разделёнными. Простой, но эффективной, стратегией является выбор в качестве порогового параметра b средней проекции объектов обучающей выборки на $\mathbf{w}$.

Линейный дискриминант Фишера

Метод ЛДФ легко обобщается на случай с несколькими классами.

При этом исходная задача распознавания классов $K_1, \ldots, K_L$ сводится к последовательности задач с двумя классами K_1' и K_2'

Зад. 1. Класс $K_1' = K_1$, класс $K_2' = \Omega \setminus K_1$

……

Зад. *L*. Класс $K_1' = K_L$, класс $K_2' = \Omega \setminus K_L$

Для каждой из *L* задач ищется оптимальное направление и пороговое правило.

Линейный дискриминант Фишера

В результате получается набор из L направлений $\mathbf{w}_1, \ldots, \mathbf{w}_L$

При распознавании нового объекта s_* по признаковому описанию $\mathbf{x}_*$ вычисляются проекции на $\mathbf{w}_1, \ldots, \mathbf{w}_L$

$$\gamma_1(\mathbf{x}_*) = \frac{(\mathbf{w}_1 \mathbf{x}_*^t)}{|\mathbf{w}_1|}, \ldots, \gamma_L(\mathbf{x}_*) = \frac{(\mathbf{w}_L \mathbf{x}_*^t)}{|\mathbf{w}_L|}$$

Распознаваемый объект относится к тому классу, соответствующему максимальной величине проекции.

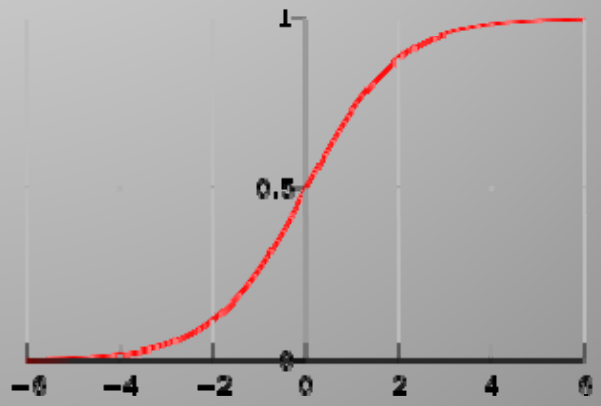
Распознавание может производится также по величинам

$$[\gamma_1(\mathbf{x}_*) - b_1], \ldots, [\gamma_L(\mathbf{x}_*) - b_L]$$

Логистическая регрессия

Целью логистической регрессии является аппроксимация плотности условных вероятностей классов в точках признакового пространства. При этом аппроксимация производится с использованием логистической функции.

$$g(z) = \frac{e^z}{e^z + 1} = \frac{1}{e^{-z} + 1}$$



Логистическая регрессия

В методе логистическая регрессия связь условной вероятности класса K с прогностическими признаками осуществляются через переменную z, которая задаётся как линейная комбинация признаков:

$$z = \beta_0 + \beta_1 X_1 + \ldots \beta_n X_n$$

Таким образом условная вероятность K в точке векторного пространства $\mathbf{x}_* = (x_{*1}, \ldots, x_{*n})$ задаётся в виде

$$\mathbf{P}(K \mid \mathbf{x}) = \frac{1}{e^{-\beta_0 - \beta_1 x_{*1} - \ldots - \beta_1 x_{*n}} + 1} = \frac{e^{\beta_0 + \beta_1 x_{*1} + \ldots + \beta_1 x_{*n}}}{e^{\beta_0 + \beta_1 x_{*1} + \ldots + \beta_1 x_{*n}} + 1} \quad (1)$$

Логистическая регрессия

Оценки регрессионных параметров $\beta_0, \beta_1, \ldots, \beta_n$ могут быть выяислены по обучающей выборке с помощью различных вариантов метода максимального правдоподобия.

Предположим, что объекты обучающей выборки сосредоточены в точках признакового пространства $\{\mathbf{x}_1, \ldots, \mathbf{x}_r\}$. При этом распределение объектов обучающей выборка по точкам задаётся с помощью набора пар $\{(m_1, k_1), \ldots, (m_r, k_r)\}$, где m_i - общее число объектов в точке $\mathbf{x}_i$, k_i - число объектов класса K в точке $\mathbf{x}_i$

Логистическая регрессия

Вероятность данной конфигурации подчиняется распределению Бернулли. Введём обозначение $p(\mathbf{x}) = \mathbf{P}(K \mid \mathbf{x})$.Функция правдоподобия может быть записана в виде

$$L(\boldsymbol{\beta}, \mathbf{x}) = \prod_{i=1}^{r} C_{m_i}^{k_i} [p(\mathbf{x})]^{k_i} [1 - p(\mathbf{x})]^{m_i - k_i}$$

Принимая во внимание что $\frac{p(\mathbf{x})}{1 - p(\mathbf{x})} = e^{\beta_0 + \beta_1 x_{*1} + \ldots + \beta_1 x_{*n}}$

$$1 - p(\mathbf{x}) = \frac{1}{1 + e^{\beta_0 + \beta_1 x_{*1} + \ldots + \beta_1 x_{*n}}}$$

Логистическая регрессия

- Получаем

$$L(\boldsymbol{\beta}, \mathbf{x}) = \prod_{i=1}^{r} C_{n_i}^{k_i} e^{k_i(\beta_0 + \beta_1 x_{*1} + \ldots + \beta_1 x_{*n})} \left(\frac{1}{1 + e^{\beta_0 + \beta_1 x_{*1} + \ldots + \beta_1 x_{*n}}}\right)^{m_i}$$

$$\ln L(\boldsymbol{\beta}, \mathbf{x}) = \sum_{i=1}^{r} \ln(C_{n_i}^{k_i}) + \sum_{i=1}^{r} k_i(\beta_0 + \beta_1 x_{*1} + \ldots + \beta_1 x_{*n}) +$$

$$+ \sum_{i=1}^{r} m_i \ln\{1/[1 + e^{(\beta_0 + \beta_1 x_{*1} + \ldots + \beta_1 x_{*n})}]\}$$

Метод k-ближайших соседей

- Простым, но достаточно эффективным подходом к решению задач распознавания является метод k-ближайших соседей. Оценка условных вероятностей $\mathbf{P}(K_i \mid \mathbf{x})$ ведётся по ближайшей окрестности V_k точки $\mathbf{x}$, содержащей k признаковых описаний объектов обучающей выборки. В качестве оценки $\mathbf{P}(K_i \mid \mathbf{x})$ выступает отношение $\frac{k_i}{k}$, где k_i - число признаковых описаний объектов обучающей выборки из K_i внутри V_k.

Метод k-ближайших соседей

Окрестность V_k задаётся с помощью функции расстояния $\rho(\mathbf{x}', \mathbf{x}'')$ заданной на декартовом произведении $\tilde{\mathbf{X}} \times \tilde{\mathbf{X}}$, где $\tilde{\mathbf{X}}$ - область допустимых значений признаковых описаний. В качестве функции расстояния может быть использована стандартная эвклидова метрика

$$\rho_e(\mathbf{x}', \mathbf{x}'') = \sqrt{\sum_{i=1}^{n} \frac{1}{n} (x_i' - x_i'')^2}$$

Метод k-ближайших соседей

Для задач с бинарными признаками в качестве функции расстояния может быть использована метрика Хэмминга, равная числу совпадающих позиций в двух сравниваемых признаковых описаниях.

Окрестность V_k ищется путём поиска в обучающей выборке k векторных описаний, ближайших в смысле выбранной функции расстояний, к описанию $\mathbf{x}^*$ распознаваемого объекта s^* .

Метод k-ближайших соседей

Единственным параметром, который может быть использован для настройки (обучения) алгоритмов в методе k – ближайших соседей является собственно само число ближайших соседей.

Для оптимизации параметра k обычно используется метод, основанный на скользящем контроле. Оценка точности распознавания производится по обучающей выборке при различных k и выбирается значение данного параметра, при котором полученная точность максимальна.

Распознавание при заданной точности распознавания некоторых классов

- Байесовский классификатор обеспечивает максимальную общую точность распознавания. Однако при решении конкретных практических задач потери, связанные с неправильной классификацией объектов, принадлежащих к одному из классов, значительно превышают потери, связанные с неправильной классификацией объектов других классов. Для оптимизации потерь необходимо использование методов распознавания с учётом предпочтительной точности распознавания для некоторых классов.

Распознавание при заданной точности распознавания некоторых классов

Одним из возможных подходов является фиксирование порога для точности распознавания одного из классов.

Оптимальное решающее правило в задаче распознавания с двумя классами K_1 и K_2, обеспечивающее максимальную точность распознавания K_2 при фиксированной точности распознавания K_1, описывается кртерием Неймана-Пирсона.

Распознавание при заданной точности для некоторых классов

Критерий Неймана-Пирсона

Максимальная точность распознавания K_2 при точности распознавания K_1 равной α обеспечивается правилом: Объект с описанием $\mathbf{x}$ относится в класс K_1, если $\mathbf{P}(K_1 \mid \mathbf{x}) \geq \eta \mathbf{P}(K_2 \mid \mathbf{x})$, где параметр η определяется из условия $\int_{\Theta} \mathbf{P}(K_1 \mid \mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \alpha$, а

$$\Theta = \{\mathbf{x} \mid \mathbf{P}(K_1 \mid \mathbf{x}) \geq \eta \mathbf{P}(K_2 \mid \mathbf{x})\}$$

$f(\mathbf{x})$ - плотность распределения $K_2 \cup K_1$ в точке $\mathbf{x}$

Распознавание при заданной точности для отдельных классов

Критерий Неймана-Пирсона может быть использован, если известны плотности распределения распознаваемых классов. Плотности могут быть восстановлены в рамках Байесовских методов обучения на основе гипотез о виде распределений.

Однако существуют эффективные средства регулирования точности распознавания при предпочтительности одного из классов, которые не требуют гипотез о виде распределения.

Распознавание при заданной точности для отдельных классов

Данные средства основаны на структуре распознающего алгоритма.

Каждый алгоритма распознавания классов $K_1, \ldots, K_L$ может быть представлен как $A = R \otimes C$ последовательное выполнение распознающего оператора R и решающего правила C :

Оператор оценок вычисляет для распознаваемого объекта s вещественный оценки за классы $K_1, \ldots, K_L$

$\{\gamma(s)_1, \ldots, \gamma(s)_L\}$

Слайд 30

ПW1 Пользователь Windows; 01.03.2012

Распознавание при заданной точности для отдельных классов

Решающее правило C производит отнесение объекта s по вектору оценок $\{\gamma_1(s), \ldots, \gamma_L(s)\}$ к одному из классов.

Распространённым решающим правилом является простая процедура, относящая объект в тот класс, оценка за который максимальна.

В случае распознавания двух классов K_1 и K_2 распознаваемый объект s_* будет отнесён к классу K_1, если $\gamma_1(s) - \gamma_2(s) > 0$

и классу K_2 в противном случае.