Лекция 8
Глава 18

КЛАССИФИКАТОР БАЙЕСА

КЛАССИФИКАТОР БАЙЕСА

Пусть набор обучающих данных **D** состоит из n точек x_i в d-мерном пространстве, и пусть y_i обозначает класс для каждой точки, где $y_i \in \{c_1, \dots, c_k\}$.

Классификатор Байеса напрямую использует теорему Байеса для прогнозирования класса для нового тестового экземпляра x. Он оценивает апостериорную вероятность $P(c_i|x)$ для каждого класса c_i и выбирает класс, который имеет наибольшую вероятность. Прогнозируемый класс для x задается как:

$$\hat{y} = argmax_{c_i}\{P(c_i|x)\}$$

ТЕОРЕМА БАЙЕСА

$$\mathrm{P}(c_i|x) = \frac{\mathrm{P}(x|c_i) \cdot P(c_i)}{P(x)}$$

где $\mathrm{P}(x|c_i)$ – *вероятность правдоподобия*, определяемая как вероятность наблюдения x при условии, что истинным классом является c_i, $P(c_i)$ - *априорная вероятность* класса c_i, а $P(x)$ - вероятность наблюдения x любого из k классов, заданных как

$$P(x) = \sum_{j=1}^{k} \mathrm{P}(x|c_j) \cdot P(c_j)$$

КЛАССИФИКАТОР БАЙЕСА

Поскольку $P(x)$ фиксирован для данной точки, правило Байеса

$$\hat{y} = argmax_{c_i}\{\mathrm{P}(c_i|x)\}$$

можно переписать как

$$\hat{y} = argmax_{c_i}\{\mathrm{P}(c_i|x)\} =$$

$$argmax_{c_i}\left\{\frac{\mathrm{P}(x|c_i)\cdot P(c_i)}{P(x)}\right\} =$$

$$argmax_{c_i}\{\mathrm{P}(x|c_i)\cdot P(c_i)\}$$

Оценка априорной вероятности

Чтобы классифицировать точки, мы должны оценить вероятность и априорные вероятности непосредственно из набора обучающих данных **D**. Пусть $\mathbf{D}_i$ обозначает подмножество точек в **D**, которые помечены классом c_i:

$$\boldsymbol{D_i} = \{x_j \in D | x_j \text{ имеет класс } y_j = c_i\}$$

Пусть размер набора данных **D** задан как |**D**| = n, и пусть размер каждого подмножества $\mathbf{D}_i$, зависящего от класса, задается как $|\ \mathbf{D}_i\ | = n_i$. Априорную вероятность для класса c_i можно оценить следующим образом:

$$\hat{P}(c_i) = \frac{n_i}{n}$$

Оценка правдоподобия

Чтобы оценить вероятность $P(x|c_i)$, мы должны оценить совместную вероятность x по всем d измерениям, то есть мы должны оценить $P(x = (x_1, x_2, \dots, x_d)|c_i)$.

Предполагая, что все измерения числовые, мы можем оценить совместную вероятность x с помощью либо непараметрического, либо параметрического подхода.

В параметрическом подходе мы обычно предполагаем, что каждый класс c_i нормально распределен вокруг некоторого среднего μ_i с соответствующей ковариационной матрицей Σ_i, оба из которых оцениваются по D_i. Таким образом, для класса c_i плотность вероятности в точке x задается как

$$f_i(x) = f(x|\mu_i, \Sigma_i) = \frac{1}{\left(\sqrt{2\pi}\right)^d \sqrt{|\Sigma_i|}} \exp\left\{-\frac{(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)}{2}\right\}$$

КЛАССИФИКАТОР БАЙЕСА

Поскольку c_i характеризуется непрерывным распределением, вероятность любой данной точки должна быть равна нулю, т. е. $\mathrm{P}(x|c_i) = 0$. Однако мы можем вычислить вероятность, рассматривая небольшой интервал $\epsilon > 0$ с центром в x:

$$\mathrm{P}(x|c_i) = 2\epsilon \cdot f_i(x)$$

Тогда апостериорная вероятность определяется как

$$\mathrm{P}(c_i|x) = \frac{2\epsilon \cdot f_i(x)P(c_i)}{\sum_{j=1}^{k} 2\epsilon \cdot f_j(x)P(j)} = \frac{f_i(x)P(c_i)}{\sum_{j=1}^{k} f_j(x)P\left(c_j\right)}$$

КЛАССИФИКАТОР БАЙЕСА

Кроме того, поскольку $\sum_{j=1}^{k} f_j(x)P(c_j)$ остается фиксированным для x, мы можем предсказать класс для x, изменив уравнение следующим образом:

$$\hat{y} = argmax_{c_i}\{f_i(x)P(c_i)\}$$

Чтобы классифицировать числовую контрольную точку x, байесовский классификатор оценивает параметры с помощью выборочного среднего и выборочной ковариационной матрицы. Выборочное среднее для класса c_i можно записать как

$$\hat{\mu}_{\boldsymbol{i}} = \frac{1}{n_i} \sum_{x_j \in D_i} x_j$$

КЛАССИФИКАТОР БАЙЕСА

выборочная матрица ковариации для каждого класса может быть оценена с помощью уравнения (2.30) следующим образом

$$\hat{\Sigma}_{\boldsymbol{i}} = \frac{1}{n_i} Z_i^T Z_i$$

где Z_i - центрированная матрица данных для класса c_i, заданная как $Z_i = D_i - 1 \cdot \hat{\mu}_{\boldsymbol{i}}^{\boldsymbol{T}}$. Эти значения могут быть использованы для оценки плотности вероятности в уравнении как $\hat{f}_{\boldsymbol{i}}(x) = f\left(x \middle| \hat{\mu}_{\boldsymbol{i}}, \hat{\Sigma}_{\boldsymbol{i}}\right)$.

КЛАССИФИКАТОР БАЙЕСА

Алгоритм 18.1 показывает псевдокод байесовского классификатора. Учитывая входной набор данных **D**, метод оценивает априорную вероятность, среднее значение и матрицу ковариации для каждого класса. Для тестирования с заданной контрольной точкой x он просто возвращает класс с максимальной апостериорной вероятностью. В стоимости обучения преобладает этап вычисления ковариационной матрицы, который занимает время $O(nd^2)$.

BAYESCLASSIFIER $(\mathbf{D} = \{(\mathbf{x}_j, y_j)\}_{j=1}^{n})$:

1 **for** $i = 1, \ldots, k$ **do**
2 $\quad \mathbf{D}_i \leftarrow \{\mathbf{x}_j \mid y_j = c_i, j = 1, \ldots, n\}$ // class-specific subsets
3 $\quad n_i \leftarrow |\mathbf{D}_i|$ // cardinality
4 $\quad \hat{P}(c_i) \leftarrow n_i / n$ // prior probability
5 $\quad \hat{\boldsymbol{\mu}}_i \leftarrow \frac{1}{n_i} \sum_{\mathbf{x}_j \in \mathbf{D}_i} \mathbf{x}_j$ // mean
6 $\quad \mathbf{Z}_i \leftarrow \mathbf{D}_i - \mathbf{1}_{n_i} \hat{\boldsymbol{\mu}}_i^T$ // centered data
7 $\quad \widehat{\boldsymbol{\Sigma}}_i \leftarrow \frac{1}{n_i} \mathbf{Z}_i^T \mathbf{Z}_i$ // covariance matrix
8 **return** $\hat{P}(c_i), \hat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i$ *for all* $i = 1, \ldots, k$

TESTING (**x and** $\hat{P}(c_i), \hat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i$, **for all** $i \in [1, k]$):

9 $\hat{y} \leftarrow \arg\max_{c_i} \{ f(\mathbf{x} | \hat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i) \cdot P(c_i) \}$
10 **return** $\hat{y}$

Категориальные атрибуты

Формально пусть X_j будет категориальным атрибутом в области $dom(X_j) = \{a_{j1}, a_{j2}, \ldots, a_{jm_j}\}$, то есть атрибут X_j может принимать m_j различных категориальных значений. Каждый категориальный атрибут X_j моделируется как m_j -мерная многомерная случайная величина Бернулли X_j, которая принимает m_j различных векторных значений $e_{j1}, e_{j2}, \ldots, e_{jm_j}$, где e_{jr} - r-й стандартный базисный вектор в $\mathbb{R}^{m_j}$ и соответствует r-му значению или символу $a_{jr} \in dom(X_j)$. Весь d-мерный набор данных моделируется как векторная случайная величина $X = (X_1, X_2, \ldots, X_d)^T$.

Категориальные атрибуты

Пусть $d' = \sum_{j=1}^{d} m_j$; категориальная точка $x = (x_1, x_2, \dots, x_d)^T$ поэтому представляется как d'-мерный двоичный вектор.

$$v = \begin{pmatrix} v_1 \\ \vdots \\ v_d \end{pmatrix} = \begin{pmatrix} e_{1r_1} \\ \vdots \\ e_{dr_d} \end{pmatrix}$$

где $v_j = e_{jr_j}$ при условии, что $x_j = a_{jr_j}$ — это r_j-е значение в области определения X_j. Вероятность категориальной точки x получается из совместной функции вероятности (PMF) для векторной случайной величины X:

$$\mathrm{P}(x|c_i) = f(v|c_i) = f\left(X_1 = e_{1r_1}, \dots, X_d = e_{dr_d} \middle| c_i\right)$$

Категориальные атрибуты

Если вероятность в точке v равна нулю для одного или обоих классов, это приведет к нулевому значению апостериорной вероятности. Чтобы избежать этого можно принять *псевдосчет*, равный 1 для каждого значения, то есть предположить, что каждое значение X встречается, по крайней мере, один раз, и увеличить этот базовый счетчик, равный 1, фактическим числом появлений наблюдаемого значения v в классе c_i. Скорректированная вероятность в v тогда дается как

$$\hat{f}(v|c_i) = \frac{n_i(v) + 1}{n_i + \prod_{j=1}^{d} m_j} \quad (18.6)$$

где $\prod_{j=1}^{d} m_j$ дает количество возможных значений **X**

Проблемы

- Отсутствие достаточного количества данных для надежной оценки совместной плотности вероятности или функции вероятности, особенно для многомерных данных.
- Даже если каждый категориальный атрибут имеет только два значения, нам нужно будет оценить вероятность для 2^d значений. Однако, поскольку для v может быть не более n различных значений, большинство отсчетов будут равны нулю.
- Для решения некоторых из этих проблем мы можем использовать сокращенный набор параметров на практике.

Наивный байесовский классификатор

Наивный байесовский подход основан на простом предположении, что все атрибуты независимы. Это приводит к гораздо более простому, но удивительно эффективному классификатору на практике. Допущение о независимости подразумевает, что вероятность принадлежности может быть разложена на произведение размерных вероятностей:

$$P(\mathbf{x} \mid c_i) = P(x_1, x_2, \ldots, x_d \mid c_i) = \prod_{j=1}^{d} P\left(x_j \mid c_i\right)$$

Числовые признаки

Для числовых признаков мы делаем допущение, что каждый из них является нормально распределенным в любом классе c_i. Пусть μ_{ij} и σ_{ij}^2 обозначают среднее значение и дисперсию для признака X_j в классе c_i. Вероятность принадлежности к классу c_i для измерения X_j определяется как:

$$P(x_j \mid c_i) \propto f(x_j \mid \mu_{ij}, \sigma_{ij}^2) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp\left\{-\frac{(x_j - \mu_{ij})^2}{2\sigma_{ij}^2}\right\}$$

Числовые признаки

В данном случае, наивное допущение соответствует установлению всех ковариаций равным нулю в $\boldsymbol{\Sigma}_i$, то есть

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \sigma_{i1}^2 & 0 & \dots & 0 \\ 0 & \sigma_{i2}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \dots & \sigma_{id}^2 \end{pmatrix}$$

Это дает

$$|\boldsymbol{\Sigma}_i| = \det(\boldsymbol{\Sigma}_i) = \sigma_{i1}^2 \sigma_{i2}^2 \cdots \sigma_{id}^2 = \prod_{j=1}^{d} \sigma_{ij}^2$$

Числовые признаки

Далее получаем

$$\boldsymbol{\Sigma}_i^{-1} = \begin{pmatrix} \frac{1}{\sigma_{i1}^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_{i2}^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \dots & \frac{1}{\sigma_{id}^2} \end{pmatrix}$$

при условии, что $\sigma_{ij}^2 \neq 0$ для всех j. Окончательно,

$$(\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) = \sum_{j=1}^{d} \frac{\left(x_j - \mu_{ij}\right)^2}{\sigma_{ij}^2}$$

Числовые признаки

Подстановка дает нам

$$P(\mathbf{x} \mid c_i) = \frac{1}{(\sqrt{2\pi})^d \sqrt{\prod_{j=1}^{d} \sigma_{ij}^2}} \exp\left\{ -\sum_{j=1}^{d} \frac{\left(x_j - \mu_{ij}\right)^2}{2\sigma_{ij}^2} \right\}$$

$$= \prod_{j=1}^{d} \left(\frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp\left\{ -\frac{\left(x_j - \mu_{ij}\right)^2}{2\sigma_{ij}^2} \right\} \right)$$

$$= \prod_{j=1}^{d} P\left(x_j \mid c_i\right)$$

Числовые признаки

Наивный байесовский классификатор использует выборочное среднее $\hat{\mu}_i = (\hat{\mu}_{i1}, \ldots, \hat{\mu}_{id})^T$ и *диагональ* выборочной ковариационной матрицы

$$\hat{\boldsymbol{\Sigma}}_i = diag(\sigma_{i1}^2, \ldots, \sigma_{id}^2)$$

для каждого класса c_i. Таким образом всего $2d$ параметром должны быть оценены в соответствии с выборочным средним и выборочной дисперсией для каждого измерения X_j.

Числовые признаки

Учитывая входной набор данных $\mathbf{D}$, метод оценивает априорную вероятность и среднее значение для каждого класса. Затем он вычисляет дисперсию $\hat{\sigma}_{ij}^2$ для каждого из атрибутов X_j со всеми d дисперсиями для класса c_i, хранящегося в векторе $\hat{\sigma}_i$. Дисперсия для атрибута X_j получается путем предварительного центрирования данных для класса $\mathbf{D}_i$ через $\mathbf{Z}_i = \mathbf{D}_i - \mathbf{1} \cdot \hat{\boldsymbol{\mu}}_i^T$. Обозначим через Z_{ij} центрированные данных для класса c_i соответствующего атрибута X_j. Тогда дисперсия выражается как $\hat{\sigma} = \frac{1}{n_i} Z_{ij}^T Z_{ij}$

NAIVEBAYES $(\mathbf{D} = \{(\mathbf{x}_j, y_j)\}_{j=1}^n)$:

1 **for** $i = 1, \ldots, k$ **do**
2 $\quad \mathbf{D}_i \leftarrow \{\mathbf{x}_j \mid y_j = c_i, j = 1, \ldots, n\}$ // class-specific subsets
3 $\quad n_i \leftarrow |\mathbf{D}_i|$ // cardinality
4 $\quad \hat{P}(c_i) \leftarrow n_i / n$ // prior probability
5 $\quad \hat{\boldsymbol{\mu}}_i \leftarrow \frac{1}{n_i} \sum_{\mathbf{x}_j \in \mathbf{D}_i} \mathbf{x}_j$ // mean
6 $\quad \mathbf{Z}_i = \mathbf{D}_i - \mathbf{1} \cdot \hat{\boldsymbol{\mu}}_i^T$ // centered data for class c_i
7 $\quad$ **for** $j = 1, .., d$ **do** // class-specific variance for X_j
8 $\quad\quad \hat{\sigma}_{ij}^2 \leftarrow \frac{1}{n_i} Z_{ij}^T Z_{ij}$ // variance
9 $\quad \hat{\boldsymbol{\sigma}}_i = \left(\hat{\sigma}_{i1}^2, \ldots, \hat{\sigma}_{id}^2\right)^T$ // class-specific attribute variances
10 **return** $\hat{P}(c_i), \hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\sigma}}_i$ for all $i = 1, \ldots, k$

TESTING (**x and** $\hat{P}(c_i)$, $\hat{\boldsymbol{\mu}}_i$, $\hat{\boldsymbol{\sigma}}_i$, **for all** $i \in [1, k]$):

11 $\hat{y} \leftarrow \underset{c_i}{\arg\max} \left\{ \hat{P}(c_i) \prod_{j=1}^{d} f(x_j | \hat{\mu}_{ij}, \hat{\sigma}_{ij}^2) \right\}$
12 **return** $\hat{y}$

Категориальные признаки

Допущение о независимости приводит к упрощению совместной функции массы вероятности в уравнении, которое можно переписать следующим образом:

$$P(\mathbf{x} \mid c_i) = \prod_{j=1}^{d} P\left(x_j \mid c_i\right) = \prod_{j=1}^{d} f\left(\mathbf{X}_j = \mathbf{e}_{jr_j} \mid c_i\right)$$

Где $f\left(\mathbf{X}_j = \mathbf{e}_{jr_j} \mid c_i\right)$ – функция массы вероятности для X_j, которую можно оценить из $\mathbf{D}_i$, как показано далее:

$$\hat{f}\left(\mathbf{v}_j \mid c_i\right) = \frac{n_i\left(\mathbf{v}_j\right)}{n_i}$$

где $n_i\left(\mathbf{v}_j\right)$ – наблюдаемая частого значения $\mathbf{v}_j = \mathbf{e}_j r_j$, соответственно r_j категориальное значение a_{jrj} для атрибута X_j в классе c_i.

Как и в полном байесовском классификаторе, если количество равно нулю, мы можем использовать метод псевдоподсчета для полученной априорной вероятности. Скорректированная оценка с псевдоподсчетом выглядит следующим образом

$$\hat{f}(\mathbf{v}_j \mid c_i) = \frac{n_i(\mathbf{v}_j) + 1}{n_i + m_j}$$

где $m_j = |dom(X_j)|$. Расширить код в алгоритме 18.2 для включения категориальных атрибутов несложно

Классификатор К-ближайших соседей

Пусть **D** - обучающий набор данных, содержащий n точек $\mathbf{x}_i \in \mathbb{R}^d$, и пусть $\mathbf{D}_i$ обозначает подмножество точек в **D**, которые помечены классом c_i, где $n_i=|\mathbf{D}_i|$. Для контрольной точки $\mathbf{x} \in \mathbb{R}^d$ и K, количества рассматриваемых соседей, пусть r обозначает расстояние от **x** до K-го ближайшего соседа в **D**.

Рассмотрим d-мерный гипершар радиуса r вокруг тестовой точки x, определенный как

$$B_d(\mathbf{x}, r) = \{\mathbf{x}_i \in \mathbf{D} \mid \delta(\mathbf{x}, \mathbf{x}_i) \leq r\}$$

Здесь $\delta(\mathbf{x}, \mathbf{x}_i)$ - это расстояние между **x** и $\mathbf{x}_i$, которое обычно считается евклидовым расстоянием, т.е. $\delta(\mathbf{x}, \mathbf{x}_i) = \|\mathbf{x} - \mathbf{x}_i\|_2$. Однако можно использовать и другие метрики расстояния. Мы предполагаем, что $|B_d(\mathbf{x}, r)| = К$.

Классификатор К-ближайших соседей

Пусть K_i обозначает количество точек среди K ближайших соседей точки $\mathbf{x}$, помеченных классом c_i, т. е.

$$K_i = \{\mathbf{x}_j \in B_d(\mathbf{x}, r) \mid y_j = c_i\}$$

Плотность условной вероятности класса в точке x можно оценить как долю точек из класса ci, лежащих внутри гипершара, деленную на его объем, т. е.

$$\hat{f}(\mathbf{x} \mid c_i) = \frac{K_i / n_i}{V} = \frac{K_i}{n_i V}$$

где $V = \text{vol}(B_d(\mathbf{x}, r))$ - объем d-мерного гипершара

Классификатор К-ближайших соседей

Используя уравнение (18.4) апостериорная вероятность $P(c_i \mid \mathbf{x})$ может быть оценена как

$$P(c_i \mid \mathbf{x}) = \frac{\hat{f}(\mathbf{x} \mid c_i)\hat{P}(c_i)}{\sum_{j=1}^{k} \hat{f}(\mathbf{x} \mid c_j)\hat{P}(c_j)}$$

Однако, так как $\hat{P}(c_i) = \frac{n_i}{n}$, мы имеем

$$\hat{f}(\mathbf{x} \mid c_i)\hat{P}(c_i) = \frac{K_i}{n_i V} \cdot \frac{n_i}{n} = \frac{K_i}{nV}$$

Классификатор K-ближайших соседей

Таким образом, апостериорная вероятность представляется как

$$P(c_i \mid \mathbf{x}) = \frac{\frac{K_i}{nV}}{\sum_{j=1}^{k} \frac{K_j}{nV}} = \frac{K_i}{K}$$

Наконец, предсказанный класс для x равен

$$\hat{y} = \arg\max_{c_i} \{P(c_i \mid \mathbf{x})\} = \arg\max_{c_i} \left\{\frac{K_i}{K}\right\} = \arg\max_{c_i} \{K_i\}$$

Поскольку K фиксировано, классификатор KNN предсказывает класс $\mathbf{x}$ как класс большинства среди своих K ближайших соседей.