Задание 1. Теория

Боймель Александр

27 февраля 2017 г.

I. Наивный байес и центроидный классификатор

По условию

$$P(x^{(k)} \mid y) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} \cdot e^{\frac{-(x^{(k)} - \mu_{yk})^2}{2 \cdot \sigma^2}}$$

Так как мы рассматриваем байесовский классификатор, то $y(x) = \underset{y}{\text{argmax}} P(y \mid x)$, где $x = (x^1, \ldots, x^n), y$ - ответ (номер класса 0 или 1) на элементе x.

Применяя формулу Байеса $y(x) = \underset{y}{\text{argmax}} P(x \mid y) \cdot P(y)$, но так как по условию распределения всех классов равны, то итоговая формула принимает вид:

$$y(x) = \underset{y}{\text{argmax}} P(x \mid y)$$

Так как рассматривается наивный байес, то компоненты x описываются независимо, т.е. $P(x \mid y) = \prod_{k=1}^{n} P(x^{(k)} \mid y)$

$$\begin{aligned} y(x) = \underset{y}{\text{argmax}} P(x \mid y) = & \\ &= \underset{y}{\text{argmax}} \left(\prod_{k=1}^{n} P(x^{(k)} \mid y) \right) \\ &= \underset{y}{\text{argmax}} \left(\frac{1}{\sqrt[\frac{n}{2}]{2 \cdot \pi \cdot \sigma^2}} \cdot e^{-\frac{\sum_{k=1}^{n} (x^{(k)} - \mu_{yk})^2}{2 \cdot \sigma^2}} \right) \\ &= \underset{y}{\text{argmax}} \left(\frac{1}{\sqrt[\frac{n}{2}]{2 \cdot \pi \cdot \sigma^2}} \cdot e^{-\frac{\rho(x, \mu_y)^2}{2 \cdot \sigma^2}} \right) \\ &= \underset{y}{\text{argmin}} (\rho(x, \mu_y)) \end{aligned}$$

Выше использовалось, что $\rho(x, y) = \sqrt{\sum_{k=1}^{n}(x^{(k)} - y^{(k)})^2}$

Таким образом ответом будет $y(x) = \underset{y}{\operatorname{argmin}}(\rho(x, \mu_y))$, т.е. y, центр которого μ_y ближе всего к x.

II. ROC-AUC случайных ответов

По условию классификатор выдает 1 с вероятностью p, 0 с вероятностью $1 - p$.

Пусть выборка размера n, n_1 - элементов 1-го класса, n_0 - 2-го, $n = n_1 + n_0$.

Тогда классификатор выдаст $p \cdot n$ единиц и $(1 - p) \cdot n$ нулей.

$$TP + FP = p \cdot n, TP + FN = n_1$$

где TP - True Positive, FN - False Negative, FP - False Positive, TN - True Negative.

По определению ROC кривая строится в координатах $TPR = \frac{TP}{TP+FN} = \frac{TP}{n_1}$, $FPR = \frac{FP}{FP+TN} = \frac{FP}{n_0}$.

Так как классификатор выдает только 0 или 1, то кривая - ломаная через 3 точки: $(0, 0), (TPR, FPR)_{p,n,n_1,n_0}, (1, 1)$

Поэтому площадь в среднем (ее математическое ожидание) будет равна сумме площадей треугольника и трапеции и равна:

$$
\begin{aligned}
E(S) = E(S_1 + S_2) = \\
&= E\left(\frac{1}{2}\cdot[FPR\cdot TPR + (1+FPR)\cdot(1-TPR)]\right) \\
&= \frac{1}{2}\cdot E(1+FPR-TPR) \\
&= \frac{1}{2}\cdot E\left(1-\frac{TP}{n_1}+\frac{FP}{n_0}\right) \\
&= \frac{1}{2}\cdot E\left(1-\frac{TP\cdot n_0 - FP\cdot n_1}{n_0\cdot n_1}\right) \\
&= \frac{1}{2}\cdot E\left(1-\frac{TP\cdot n_0 - (p\cdot n - TP)\cdot n_1}{n_0\cdot n_1}\right) \\
&= \frac{1}{2}\cdot\left(1-E\left(\frac{TP\cdot n - p\cdot n\cdot n_1}{n_0\cdot n_1}\right)\right) \\
&= \frac{1}{2}\cdot\left(1-\frac{n\cdot E(TP - p\cdot n_1)}{n_0\cdot n_1}\right) \\
&= [\text{т.к. } E(TP) = E\left(\sum_{k=1}^{n_1} I(y_k=1)\right) = n_1\cdot P(y(x)=1) = p\cdot n_1] \\
&= \frac{1}{2}\cdot(1-0) \\
&= \frac{1}{2}
\end{aligned}
$$

Таким образом $E(S) = ROC - AUC = \frac{1}{2}$. Ч.Т.Д.

III. Ошибка 1NN и оптимального байесовского классификатора

Байесовский классификатор отвечает:

$$
\begin{cases}
1, & P(1\mid x) > P(0\mid x) \\
0, & P(0\mid x) > P(1\mid x)
\end{cases}
$$

Поэтому матожидание ошибки байесовского классификатора:

$$E_B(x) = \min\left[P(0\mid x), P(1\mid x)\right]$$

Для 1NN классификатора матожидание ошибки:

$$E_N(x) = P(y \neq y_n)$$

где y - настоящий класс x, y_n - класс ближайшего соседа x_n к элементу x.

$$
\begin{aligned}
E_N(x) = P(y \neq y_n) = \\
&= P(y = 0, y_n = 1) + P(y = 1, y_n = 0) \\
&= P(0 \mid x) \cdot P(1 \mid x_n) + P(1 \mid x) \cdot P(0 \mid x_n) \\
&= [\text{воспользуемся асимптотическим приближением } x_n \to x] \\
&= 2 \cdot P(0 \mid x) \cdot P(1 \mid x) \\
&= 2 \cdot (1 - E_B(x)) \cdot E_B(x) \\
&= 2 \cdot E_B(x) - 2 \cdot (E_B(x))^2 \\
&\leq 2 \cdot E_B(x)
\end{aligned}
$$

$$E_B \leq \lim_{N \to \infty} E_N \leq 2 \cdot E_B$$

Ч.Т.Д.