Машинное обучение – 1. Теоретические задачи

Каргальцев Степан

26.02.2017

1. **Наивный байес и центроидный классификатор**

   Покажите, что если в наивном байесовском классификаторе классы имеют одинаковые априорные вероятности, а плотность распределения признаков в каждом классе имеет вид $P(x^{(k)}|y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^{(k)}-\mu_{yk})^2}{2\sigma^2}}, x^(k), k = 1 \ldots n$ *— признаки объекта* x*, классификация сводится к отнесению объекта* x *к классу* y*, центр которого* $\mu_y = (\mu_1, \ldots, \mu_n)^T$ *ближе всего к* x*.*

$$\hat{y} = \underset{y}{\text{argmax}} \left[\prod_{k=1}^{n} P(x^{(k)}|y)P(y) \right] = |P(y) \text{ одинаковы}| = \underset{y}{\text{argmax}} \left[\prod_{k=1}^{n} P(x^{(k)}|y) \right] =$$

$$\underset{y}{\text{argmax}} \left[\prod_{k=1}^{n} \ln P(x^{(k)}|y) \right] = \underset{y}{\text{argmax}} \left[-\sum_{k=1}^{n} \frac{(x^{(k)} - \mu_{yk})^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2) \right] = \underset{y}{\text{argmin}} \left[\sum_{k=1}^{n} (x^{(k)} - \mu_{yk})^2 \right] =$$

$$\underset{y}{\text{argmin}}\, \rho(x, \mu_y)$$

2. **ROC-AUC случайных ответов**

   Покажите, что «треугольный ROC-AUC» (см.лекцию 2) в случае, когда классификатор дает случайные ответы – $a(x) = 1$ с вероятностью p и $a(x) = 0$ с вероятностью $1 - p$, будет в среднем равен 0.5, независимо от p и доли класса 1 в обучающей выборке.

   Если пороговое значение 0 (то есть все ответы положительны), то мы получаем TPR = 1, FPR = 1 и точку (1, 1)

   Если пороговое значение 1 (то есть все ответы отрицательны), то мы получаем TPR = 0, FPR = 0 и точку (0, 0)

   Иначе мы получаем TPR = $\frac{\sum\limits_{x \in 1} a(x)}{|1|} =: \mathcal{Y}$, FRP = $\frac{\sum\limits_{x \in 0} a(x)}{|0|} =: \mathcal{X}$.

   Площадь треугольника на этих трех точках равна $S = \frac{1}{2}(\mathcal{X}, \mathcal{Y}) \times (1, 1) = \frac{1}{2}(\mathcal{X} - \mathcal{Y}) = \frac{\sum\limits_{x \in 0} a(x)}{2|0|} - \frac{\sum\limits_{x \in 1} a(x)}{2|1|}$.

   $ES = \frac{1}{2}E\left[\frac{\sum\limits_{x \in 0} a(x)}{|0|} - \frac{\sum\limits_{x \in 1} a(x)}{|1|}\right] = \frac{\sum\limits_{x \in 0} Ea(x)}{2|0|} - \frac{\sum\limits_{x \in 1} Ea(x)}{2|1|} = \frac{1}{2}(Ea(x) - Ea(x)) = 0$

   То есть площадь под $ROC - AUC$ кривой будет в среднем равна $E(0.5 + S) = 0.5$

3. **Ошибка 1NN и оптимального байесовского классификатора**

   Утверждается, что метод одного ближайшего соседа асимптотически (при условии, что максимальное по всем точкам выборки расстояние до ближайшего соседа стремится к нулю) имеет матожидание ошибки не более чем вдвое больше по сравнению с оптимальным байесовским классификатором (который это матожидание минимизирует).

   Покажите это, рассмотрев задачу бинарной классификации. Достаточно рассмотреть вероятность ошибки на фиксированном объекте x*, т.к. матожидание ошибок на выборке размера* V *будет просто произведением* V *на эту вероятность. Байесовский классификатор ошибается на объекте* x *с вероятностью:*

$$E_B = \min\{P(1|x), P(0|x)\}$$

Условные вероятности будем считать непрерывными функциями от $x \in \mathbb{R}^m$ *, чтобы иметь возможность делать предельные переходы. Метод ближайшего соседа ошибается с вероятностью:*

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

Здесь y *— настоящий класс* x*, а* y_n *— класс ближайшего соседа* x_n *к объекту* x *в предположении, что в обучающей выборке* n *объектов, равномерно заполняющих пространство.*

Докажите исходное утверждение, выписав выражение для E_N *(принадлежность к классам* 0 *и* 1 *для объектов* x *и* x_n *считать независимыми событиями) и осуществив предельный переход по* n*.*

$$E_N = P(1|x_n) \cdot P(0|x) + P(0|x_n) \cdot P(1|x) \to 2P(0|x)P(1|x) =$$

$$2\min\{P(0|x), P(1|x)\} \cdot \max\{P(0|x), P(1|x)\} \leqslant 2\min\{P(0|x), P(1|x)\} = 2E_B$$