Теоретические задачи

4.1 Наивный байес и центроидный классификатор

Наивный байесовский классификатор относит объект к классу по правилу $a(x) = \arg\max\limits_y P(y) \cdot \prod_{k=1}^{n} P(x^{(k)}|y)$. Рассмотрим наивный байесовский классификатор, в котором классы имеют одинаковое априорное распределение P(y), а плотность распределения признаков в каждом классе имеет вид $P(x^{(k)}|y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^{(k)}-\mu_{yk})^2}{2\sigma^2}}$. Тогда P(y) постоянно при любом y, значит его можно убрать из-под аргумента. Распишем a(x):

$$a(x) = \arg\max_y P(y) \cdot \prod_{k=1}^{n} P(x^{(k)}|y) = \arg\max_y \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^{(k)}-\mu_{yk})^2}{2\sigma^2}} =$$

$$= \arg\max_y \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x^{(k)}-\mu_{yk})^2} = \arg\min_y \sum_{k=1}^{n} (x^{(k)} - \mu_{yk})^2$$

, где $\sum_{k=1}^{n}(x^{(k)} - \mu_{yk})^2$ - квадрат расстояния от точки x до центра класса y μ_y. Значит такой наивный байесовский классификатор относит точку к тому классу, чей центр ближе всего к ней.

4.2 ROC-AUC случайных ответов

Для того чтобы посчитать ROC-AUC, необходимо построить ROC-кривую - график зависимости TPR от FPR, где $TPR = \frac{TP}{TP+FN}$, $FPR = \frac{FP}{FP+TN}$ - и измерить площадь под построенной кривой. Обозначим общее число объектов - N, число объектов класса 1 - n. Тогда число объектов класса 0 - (N - n). Найдём величины TP, FP, TN и FN в случае, когда классификатор даёт ответ 1 с вероятностью p и ответ 0 с вероятностью (1 - p). TP(True Positive) равно числу элементов класса 1, которые классификатор отнёс к классу 1. Значит,

$$TP = n \cdot p$$

. Аналогично выражаем FP, TN, FN:

$$FP = (N - n) \cdot p$$

$$TN = (N - n) \cdot (1 - p)$$

$$FN = n \cdot (1 - p)$$

Найдём TPR и FPR:

$$TPR = \frac{TP}{TP + FN} = \frac{n \cdot p}{n \cdot p + n \cdot (1 - p)} = \frac{n \cdot p}{n} = p$$

$$FPR = \frac{FP}{FP + TN} = \frac{(N - n) \cdot p}{(N - n) \cdot p + (N - n) \cdot (1 - p)} = p$$

Получается, что в данном случае TPR = FPR. Тогда ROC-кривая - прямая соединяющая точки (0, 0) и (1, 1), площадь под которой равна 0.5.

4.3 Ошибка 1NN и оптимального байесовского классификатора

Рассмотрим вероятность ошибки на фиксированном объекте x. Для байесовского классификатора вероятность ошибки - $E_B = \min\{P(1|x), P(0|x)\}$. Найдём вероятность ошибки метода ближайшего соседа:

$$E_N = P(y \neq y_n) = P(y = 0) \cdot P(y_n = 1) + P(y = 1) \cdot P(y_n = 0) =$$

$$= P(0|x) \cdot P(1|x_n) + P(1|x) \cdot P(0|x_n) \underset{n\to\infty}{\longrightarrow} 2 \cdot P(0|x) \cdot P(1|x) = 2 \cdot E_B \cdot (1 - E_B) \leq 2 \cdot E_B$$

Использовано то, что при увеличении числа точек ближайший сосед приближается к x и распределение классов на ближайших соседях стремится к распределению классов на всей выборке, а также то, что $P(1|x) + P(0|x) = 1$ и одна из этих вероятностей равна E_B, значит, другая равна $(1 - E_B)$.