theory

28 февраля 2017 г.

1 Задача 1

Байесовский классификатор ищет *argmax* выражения $P(x|y)P(y)$. Поскольку $P(y) - const$, можно рассматривать только $P(x|y)$. Обозначим за $\rho(x,y)$ декартову метрику в пространстве признаков и запишем совместную условную плотность:

$P(x|y) = \prod_{i=1}^{n} P(x^{(i)}|y) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n exp(-\frac{\sum(x^{(i)}-\mu_{yi})^2}{2\sigma^2}) = const \cdot exp(-\sum(x^{(i)} - \mu_{yi})^2) = const \cdot exp(-\rho(x,\mu_y)^2)$

Из формулы видно, что значение максимально при минимальном $\rho(x,\mu_y)$, то есть байес выберет класс, центр которого будет ближе к x.

2 Задача 2

Пусть q - доля класса 1. Обозначим случайную величину TPR за y, FPR за x.

Посчитаем матожидания x и y(можно считать, что объект один). $Ey = E(pred = 1|class = 1) = E(pred = 1) = p$. $Ex = E(pred = 1|class = 0) = E(pred = 1) = p$. Выразим AUC через x и y. AUC - площадь под графиком $(0,0),(x,y),(1,1)$. $AUC = \frac{xy}{2} + \frac{(1-x)(y+1)}{2} = 0.5 + x - y$. $E\ AUC = 0.5 + p - p = 0.5$. Это верно при любых p и q.

3 Задача 3

$E_N = P(y_n \neq y) = P(y = 0)P(y_n = 1) + P(y = 1)P(y_n = 0)$, т.к. принадлежности x и x_n к классам независимы.

$P(y = 0)P(y_n = 1) = P(0|x)P(1|x_n)$ из оптимальности байесовского классификатора.

В пределе по n $P(1|x_n) \to P(1|x)$ из непрерывности условных вероятностей и так как $\rho(x_n,x) \to 0$.

Тогда в пределе $P(y = 0)P(y_n = 1) \leq P(0|x)$ и $P(y = 0)P(y_n = 1) \leq P(1|x)$. То есть $P(y = 0)P(y_n = 1) \leq E_B$.

Аналогично для $P(y = 1)P(y_n = 0)$. Итого, $E_N \leq 2E_B$.