

Наивный баиес и центроидный классификатор

По определению байесовского классификатора, необходимо найти

$$class(x) = \underset{y}{argmax} P(y) \prod_{k=1}^n P(x^{(k)}|y)$$

Априорная вероятность одинакова для всех классов и равна P_c . Плотность распределения признаков равна

$$P(x^{(k)}|y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^{(k)} - \mu_{yk})^2}{2\sigma^2}} \quad \forall k \in \{1, n\}$$

Тогда:

$$class(x) = \underset{y}{argmax} P_c \prod_{k=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x^{(k)} - \mu_{yk})^2}{2\sigma^2}}$$

Прологарифмируем рассматриваемое произведение:

$$L(x, y) = \ln(P_c) + \ln\left(\frac{n}{\sqrt{2\pi\sigma^2}}\right) + \sum_{k=1}^n -\frac{(x^{(k)} - \mu_{yk})^2}{2\sigma^2}$$

Нахождение $\underset{y}{argmax} (L(x, y))$ эквивалентно нахождению $\underset{y}{argmax}$ исходной функции. Рассмотрим только ту часть функции, которая зависит от y (остальная часть не влияет на $\underset{y}{argmax} (L(x, y))$) Остается только

$$\sum_{k=1}^n -(x^{(k)} - \mu_{yk})^2 = -\rho^2(x, \mu_y)$$

То есть нахождение $\underset{y}{argmax} L(x, y)$ эквивалентно минимизации расстояния от x до μ_y по всевозможным y . Продолжая цепочку эквивалентностей обратно, получаем требуемое утверждение.

ROC-AUC случайных ответов

Пусть в выборке N элементов, α - доля объектов класса 0. Соответственно доля объектов класса 1 будет равна $1 - \alpha$. Покажем, что все зависимости от N , α и p , True Positive Rate в среднем будет равен False Positive Rate.

$$TPR = \frac{TP}{TP + FN} = \frac{p\alpha N}{p\alpha N + (1 - p)\alpha N} = \frac{p}{p + 1 - p} = p$$

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

Следовательно в среднем ROC кривая будет отрезок $(0, 0)$ - $(1, 1)$. Поэтому площадь под ней будет равняться в среднем $ROCAUC = 0.5$.