4 Теоретические задачи

4.1 наивный байес и центроидный классификатор

$$\underset{y}{argmax}(\prod_{k=1}^{n}\frac{1}{2\pi\sigma^2}e^{\frac{(x^{(k)}-\mu_{y_k})^2}{2\sigma^2}}Pr(y)) =$$

$$= \underset{y}{argmin}(\sum_{k=1}^{n}(x^{(k)}-\mu_{y_k})^2)$$

Посколько $Pr(y) = const$, а максимум оставшейся функции является минимумом аргумента экспоненты, в которой также можно избавиться от константного множителя

4.2 ROC-AUC случайных ответов Пусть количество 1 в выборке равно k, размер выборки n, тогда:

$$\xi_i = Bern(p)$$

$$tp = \sum_{i=1}^{k}\xi_i, fp = \sum_{i=1}^{n-k}\xi_i$$

$$tn = \sum_{i=1}^{n-k}1-\xi_i, fn = \sum_{i=1}^{k}1-\xi_i$$

Получаем искомые случайные величины:

$$FPR = \frac{fp}{fp+tn} = \frac{\sum_{i=1}^{n-k}\xi_i}{\sum_{i=1}^{n-k}\xi_i + \sum_{i=1}^{n-k}(1-\xi_i)} = \frac{\sum_{i=1}^{n-k}\xi_i}{n-k}$$

$$TPR = \frac{tp}{tp+fn} = \frac{\sum_{i=1}^{k}\xi_i}{\sum_{i=1}^{k}\xi_i + \sum_{i=1}^{k}(1-\xi_i)} = \frac{\sum_{i=1}^{k}\xi_i}{k}$$

взяв математическое ожидание, получим:

$$\mathbb{E}FPR = p$$

$$\mathbb{E}TPR = p$$

Значит средняя точка лежит на диагонали квадрата. Значит площадь в среднем равна $\frac{1}{2}$

4.3 Ошибка 1NN оптимального байесовского классификатора

$$\lim_{n\to\infty}Pr(y\neq y_n) = \lim_{n\to\infty}\sum_{y\neq y_n\in\{0,1\}}Pr(y,y_n|x,x_n) =$$

$$= \lim_{n\to\infty}\sum_{y\neq y_n\in\{0,1\}}Pr(y|x)Pr(y_n|x_n) = \sum_{y\in\{0,1\}}Pr(y|x)(1-Pr(y|x)) =$$

$$= Pr(0|x)(1-Pr(0|x)) + Pr(1|x)(1-Pr(1|x)) =$$

$$= 2\max_{y\in\{0,1\}}Pr(y|x)(1-\max_{y\in\{0,1\}}Pr(y|x)) \leq 2E_B =$$

$$2(1-\max_{y\in\{0,1\}}Pr(y|x))$$