№35

Задача восстановления регрессии

Функция потерь $L(y', y) = (y'-y)^2$

Доказать, что если $f^*(x) = \underset{c}{\operatorname{argmin}}\, E((Y-c)^2 | X=x)$, то

$f^*(x) = E(Y|X=x)$

1. Разложим $(Y-c)^2$ как $Y^2 - 2Yc + c^2$, тогда

$$f^*(x) = \underset{c}{\operatorname{argmin}}\, E(Y^2 - 2Yc + c^2 / X=x)$$

Т.к. мат. ожидание линейно, то

$$E(Y^2 - 2Yc + c^2 / X=x) = E(Y^2 | X=x) - 2E(Yc|X=x) +$$
$$+ E(c^2|X=x) = E(Y^2|X=x) - 2cE(Y|X=x) + c^2$$

$E(Y^2|X=x)$ – не зависит от c $\Rightarrow$

$f^*(x)$ – будет принимать мин. знач. в случае, если

$$\frac{d}{dc}\left(-2cE(Y|X=x) + c^2\right) = 0$$

$$-2E(Y|X=x) + 2c = 0 \Rightarrow$$

$$c = E(Y|X=x)$$

$$f^*(x) = E(Y|X=x)$$

2. ~~$R(f^*) = \int_x \int_y L(f^*(x), y)) p(x.$~~

2 $R(f^*) = \int_x \int_y L(f^*(x), y)\, p(x,y)\, dx dy =$

$\int_x \int_y L(f^*(x), y)\, p(y|x)\, dy\, p(x)\, dx$, где $\int_y L(f^*(x), y) p(y|x) dy$ – мат. ожидание – $E(L(f^*(x), Y)|X=x) \Rightarrow$

$$R(f^*) = \int_x E(L(f^*(x), Y)|X=x)\, p(x)\, dx$$

$$L(f^*(x), Y) = (f^*(x) - Y)^2 \Rightarrow$$

$$R(f^*) = \int_x E((f^*(x) - Y)^2 | X = x) p(x) dx$$

т.к $f^*(x) = E(Y | X = x)$, то

$$R(f^*) = \int_x E(E(Y | X = x) - Y)^2 | X = x) p(x) dx$$

Представим

$$E(E(Y|X = x) - Y)^2 | X = x) \text{ как}$$

$$D_Y(Y | X = x) \quad \text{т.к} \quad D(x) = E((X - E(x))^2)$$

$$R(f^*) = \int_x D_Y(Y | X = x) p(x) dx$$

По определению мат. ожидания

$$E(x) = \int_x X p(x) dx \Rightarrow$$

$$R(f^*) = E_x(D_Y(Y | X = x))$$

№36

Задача восстановление регрессии

$L(y', y) = |y' - y|$

Доказать $\min R(f)$ при $f(x) = \text{median}(Y|X=x)$

$R(f) = \int_x \int_y L(f(x), y)\, p(y|x)\, dy\, p(x)\, dx$

Запишем $\int_y L(f(x), y)\, p(y|x)\, dy$ как

$E(L(f(x), Y)|X=x) = E(|f(x) - Y| \,|X=x)$ по условию $\Rightarrow$

$R(f) = \underset{f(x)}{\text{argmin}}\, E(|f(x) - Y| \,|X=x) =$

$\ominus\ \underset{c}{\text{argmin}}\, E(|c - Y| \,|X=x)$

По определению мат. ожид. ~~через~~ ~~случайного~~ ~~случайной~~ ~~величины~~ для абсолют. непрерыв. случ. вел.

$E|x| = \int_{-\infty}^{\infty} x f(x)\, dx$, тогда

$E(|c - Y| \,|X=x) = \int_{-\infty}^{\infty} |c - Y| f(x)\, dx$

Запишем интеграл как сумму интегралов

$\int_{-\infty}^{c} (c - Y) f(x)\, dx + \int_{c}^{\infty} (c - Y) f(x)\, dx = \int_{-\infty}^{c} c f(x)\, dx - \int_{-\infty}^{c} Y f(x)\, dx +$

$\int_{c}^{\infty} c f(x)\, dx - \int_{c}^{\infty} Y f(x)\, dx$

т.к ищем мин. знач

$\frac{dR(f)}{dc} = \int_{-\infty}^{c} f(x)\, dx - \int_{c}^{\infty} f(x)\, dx = 0$

получаем систему уравнений

$$\begin{cases} \int_{-\infty}^{c} f(x)\, dx - \int_{c}^{\infty} f(x)\, dx = 0 \\ \int_{-\infty}^{\infty} f(x)\, dx = 1 \end{cases}$$

Из которой следует, что

$\int_{-\infty}^{c} f(x)dx$ и $\int_{c}^{\infty} f(x)dx = \frac{1}{2}$, что соответствует

значению $f(x) = median(Y|X=x)$

Что и требовалось доказать

№38 Задача классификации

$K = 2 \quad \{0, 1\}$

$L(y', y) = \begin{array}{c|c} 0 & l_0 \\ \hline l_1 & 0 \end{array}$, где $L(0,0) = L(1,1) = 0$
$L(1,0) = l_1,\ L(0,1) = l_0$

Доказать, что $f(x) = \underset{y \in \{0,1\}}{\operatorname{argmax}}\, l_y \Pr(y|x)$

$$\Pr(y=0|X=x) = \frac{\Pr(X=x|y=0)\Pr(y=0)}{\Pr(X=x)}$$

$$\Pr(y=1|X=x) = \frac{\Pr(X=x|y=1)\Pr(y=1)}{\Pr(X=x)}$$

Вероятность ошибки можно записать как

$$R(f^*) = \sum_{y=0}^{K} l_y \Pr(y=y|X=x)$$

$$f^*(x) = \operatorname{argmin} R(f) = \operatorname{argmin} \sum_{y=0}^{K} l_y \Pr(y=y|X=x)$$

Если $l_0 < l_1$, то $R(f)$ будет минимальна при $y=1$

Если $l_0 > l_1$, то $R(f)$ будет минимальна при $y=0$

~~Если $l_0 = l_1$, ???~~

В таком случае $f^*(x) = \underset{y \in \{0,1\}}{\operatorname{argmax}}\, l_y \Pr(y|x)$

будет выбирать класс соответствующий произведению минимального знач. функции ошибки на апостер. вер.

№37

$$f^*(x) = moda(Y|X=x)$$

Найти $L(y', y)$

Чтобы минимум $R(f)$ давала мода, необходимо чтобы функция потерь имела вид:

$$L(y', y) = \begin{cases} 0, \text{ если } y' = moda(Y|X=x) \\ 1, \text{ если } y' \neq moda(Y|X=x) \end{cases}$$

т.е имела индикаторный вид

В таком случае апроксимирум средний риск для рассмотрения приблимсета

$$R(f) = \hat{R}(f) = \frac{1}{N}\sum_{i=1}^{N} L(f(x^i), y^i)$$

В случае если $f(x^i)$ равно моде, то ~~средние сумма~~ значение 0 будет чаще встречаться в последовательности, в виду того, что $f(x^i)$ соответствую наиболее частому значению, из-за которого функция потерь будет принимать значение 0

39 Задача классификации с K классами

$$L(y',y) = l_{y'y} \quad (y',y = 1,2 \dots K)$$

$$L_{y'y} = \begin{pmatrix} L_{11} & L_{21} & \cdots & L_{K1} \\ L_{12} & L_{22} & \cdots & \vdots \\ \vdots & \cdots & L_{33} & \vdots \\ L_{1K} & \cdots & \cdots & L_{KK} \end{pmatrix}$$

В таком случае байесовская ошибка имеет вид:

$$R(f) = \int_x \sum_{y=1}^{K} L_{y'y} P_r(y|x) p(x)\,dx \;; R(f) \to \min$$

$$\Rightarrow \quad f^*(x) = \underset{y' \in \{1 \dots K\}}{\operatorname{argmin}} \sum_{y=1}^{K} L_{y'y} P_r(y|x)$$