Билет 1

Понятие точечной оценки параметра. Состоятельность и несмещённость оценок

Точечная оценка Пусть $x_1, \dots x_n$ - набор измерений какой-то величины. $\xi_1, \xi_2, \dots \xi_n$ - набор iid. а - неизв. параметр.

ЦПТ: $\xi_1, \dots \xi_n$ - iid, $M\xi_i = a$, $D\xi_i = \sigma^2 \Rightarrow$ при $n \to \infty$ $\left(\frac{\xi_1 + \dots + \xi_n - na}{\sigma\sqrt{n}} < x\right) \to$ по распр. $\to N(0,1)$ или $\frac{\xi_1 + \xi_n - na}{\sigma\sqrt{n}} \to N(0,1)$

Пусть $F(x)$ - функция распр. ξ_k, предположим, она зав. от Θ

Опр Оценкой пар. Θ наз. любая функция от $(x_1, \dots x_n)$ (при этом хотим, чтобы знач. этой функции были как можно ближе к истинным)

Изначально не знаем распр. ξ_i, хотим найти величину, от которой она зависит.

Опр: Θ_n^* - несмещённая оценка Θ, если $E\Theta_n^* = \Theta$

Опр.: Θ_n^* - состоятельная оценка, если $\forall \varepsilon > 0$ $P(|\Theta_n^* - \Theta| < \varepsilon) \to 1$ при $n \to \infty$

Пример: $\xi_1, \dots \xi_n$ - iid, $E\xi_i$ - ?

$\Theta_n^* = \frac{\xi_1 + \dots + \xi_n}{n}$; $\Theta_n^* = \xi_i$ - несмещённые оценки

Опр. Θ_n^* - эффективна в ин. множестве оценок, если она обладает в нём минимальной дисперсией

Билет 2

Метод максимального правдоподобия (ММП). Его оценки параметров

Основной способ получения (точечных) оценок по выборке - метод макс. правдоподобия

Опр. Ф-ция правдоподобия $L = L(x_1, \dots x_n) = p(x_1, \dots x_n, \Theta)$ совм. вер. выборки (какой?)

Если iid: $L = P(x_1, \Theta) P(x_2, \Theta) \dots P(x_n, \Theta)$ $x_1, \dots x_n$ - фикс., Θ могут быть различными для разных x_i

Опр. Оценка максимального правдоподобия $\Theta_{м.п.} = \arg\max_{\Theta} L$

Часто удобнее искать $\max \ln L(\Theta)$, которая совп. с $\max L(\Theta)$ в силу монотонности логарифма

$\frac{\partial L}{\partial \Theta} = 0 \Rightarrow \max \frac{\partial \ln L}{\partial \Theta} = 0$

Пример Бернулли x_k: $\begin{cases} 0, p \\ 1, 1-p \end{cases}$

$L(x_1, \dots x_n, p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$

$\ln L = \sum_{i=1}^{n} x_i \ln p + (1 - x_i) \ln(1-p)$

$\frac{\partial \ln L}{\partial p} = \sum_{i=1}^{n} \frac{x_i}{p} - \sum (1 - x_i) \frac{1}{1-p}$

На самом деле $L = P_{\xi_1(\Theta), \xi_2(\Theta), \dots \xi_n(\Theta)}(x_1, \dots x_n)$, где $\xi_i(\Theta)$ описывает зависимость случайной величины ξ от Θ, x_i - измерения

Билет 3

Интервальные оценки. Доверительный интервал

…вместо точечной оценки Θ^* можно указывать интервал $(\underline{\Theta}, \overline{\Theta})$. $\underline{\Theta}(x_1, \dots x_n)$, $\overline{\Theta}(x_1, \dots x_n)$: $P(\underline{\Theta}(x_1, \dots x_n) < \Theta < \overline{\Theta}(x_1, \dots x_n)) = 1-2\alpha$

$(\underline{\Theta}, \overline{\Theta})$ - доверительный интервал для Θ с дов. вероятностью $1-2\alpha$.

…Пример: Пусть $\overline{x} = \frac{x_1 + \dots + x_n}{n}$, $x_1, \dots x_n \sim N(a, \sigma)$ iid, $E\overline{x} = a$, $D\overline{x} = \frac{\sigma^2}{n}$

…Центральной предельной теореме $\frac{\overline{x}-a}{\sigma/\sqrt{n}} \sim N(0,1)$

…сли $\alpha = \frac{1}{\sqrt{2\pi}} \int_{u_\alpha} e^{-x^2/2} dx$ (возьмем такой u_α)

…огда дов. интервал с дов. вер-ностью $1-2\alpha$: $(\overline{x} - u_\alpha \frac{\sigma}{\sqrt{n}}, \overline{x} + u_\alpha \frac{\sigma}{\sqrt{n}})$

Если для … посл. значений пар. $\underline{\Theta}_i, \overline{\Theta}_i$ $(i = 1, \dots n)$

$P(\underline{\Theta}(x_1, \dots x_n) < \Theta < \overline{\Theta}(x_1, \dots x_n)) \xrightarrow{n\to\infty} 1-2\alpha$, то такой интервал наз.

…асимптотическим.

Билет 4

Подход к проверке гипотезы о законе распределения. Статистические критерии. Статистика χ^2 и расп. Теорема о статистике χ^2 (без доказательства)

… Статистическая гипотеза - некоторое предположение о

…ах и хар-ках иссл. ген. совокупностей

… проверяемые на осн. анализа наблюдаемых выборок. При

…ом отсутствие опроверж. не док-вает, что гипотеза верна.

…сть $x_1, \dots x_n$ - iid, $F(x)$ - некоторая данная ф-ция

…вляется ли F ф-цией распр. x? и разобьем числовую ось на про-

…непересекающихся случаем промежутки: $(-\infty, z_1), (z_1, z_2) \dots (z_{j-1}, z_j), (z_j, +\infty)$ (1)

…ли F ф-ция распр. x_i, то $p_1 = P(x_1 \in (-\infty, z_1)) = F(z_1)$

$p_2 = P(x_2 \in (z_1, z_2)) = F(z_2) - F(z_1)$

$\vdots$

$p_{i-1} = P(x_{i-1} \in (z_{j-1}, z_j)) = F(z_j) - F(z_{j-1})$

$p_i = P(x_i \in (z_j, +\infty)) = 1 - F(z_j)$

… случ. вел. x_i сводится полиномиальная ~~…~~ схема с

…испытаниями след. образом: i-ое испытание - попадание

… x_i в некоторый интервал из (1)

…усть $n_i = m_i(x_1, \dots x_n)$ - число знач., попавших в (z_i, z_{i+1})

… $= np_i$, $\eta_{n,j} = \sum_{i=1}^{j} \frac{(m_i - np_i)^2}{np_i}$ - величина отклонения

… большие величины, то ф-ция распр. далека

… $\eta_{n,j} > c$, где c - … и гипотеза против. наблюд.

… наблюд. значений

… добиться $P(\eta_{n,j} > c) = \alpha$, где α мала

… Пусть случ. вел. $x_1, \dots x_n$ расп. по закону $N(0,1)$ и iid. Тогда

…опред. с.в. $R_n = x_1^2 + \dots + x_n^2$ наз. распр. χ^2 с n степенями свободы $R_n^2 \sim \chi_n^2$

Теорема $\forall x \in [0,1]$ при $n \to \infty$ $P(\eta_{n,s} < x) \to P(\chi^2_{n-1} < x)$, где

с.в. χ^2_{n-1}, имеет χ^2-распр. с $n-1$ ст. своб.

Опр. Статистический критерий: Множества отвержения, если $\eta_{n,s} \geq C$

Множество не противоречит набл., если принадлежит проц. проц. дополнение события.

Билет 5

Выбор из двух гипотез. Ошибки первого и второго рода

Пусть $\xi_1, \dots \xi_n \sim N(a, \sigma)$, i.i.d, σ - известен, a - неизв.

2 гипотезы: $H_0: a = a_0$, $H_1: a = a_1$, $a_0 < a_1$, H_0 - осн. гипотеза, H_1 - конкурирующая.

Ф-ции распред. величин ξ для разных гипотез различны, $P_0(P_1)$ - вер-ность события, вычисл. при гипотезе $H_0(H_1)$

Выберем оценку $\bar{\xi} = \frac{\xi_1 + \dots + \xi_n}{n}$ (она несмещ. и состоятельная)

Критерий: Выберем $a_0 < C < a_1$. Если $\bar{\xi} > C$, примем H_1, иначе H_0

Опр. Ошибка первого рода - вероятность принять H_1, когда верна H_0.

$\alpha = P_0(\xi > C)$

Опр. Ошибка второго рода - вероятность принять H_0, когда верна H_1.

$\beta = P_1(\xi < C)$

	Верная гипотеза H_0	Верная гипотеза H_1
Принимаем H_0	H_0 +	H_0 − β
Принимаем H_1	H_0 − α	H_0 +

Пусть $\alpha = P_0(\bar{\xi} > C)$ задано. Тогда $\alpha = P_0(\xi > C) = P_0\left(\frac{\bar{\xi} - a_0}{\sigma/\sqrt{n}} > (C - a_0)\frac{\sqrt{n}}{\sigma}\right)$

Следовательно, $C = a_0 + u_\alpha \frac{\sigma}{\sqrt{n}}$, т.к. $\sqrt{n}(\bar{\xi} - a_0)/\sigma$ распр. норм. и станд.

Аналог. для ошибки второго рода $\beta = P_1(\bar{\xi} < C)$

Получаем $u_\alpha + u_\beta = \frac{a_1 - a_0}{\sigma}\sqrt{n} \Rightarrow$ можно выбрать α, β в завис. от ситуации неотрицательных сверх. с какой ошибок решений

Билет 6

Вероятностная постановка задачи классификации

Опр. Пусть с.в. ξ, η такова, что с.в. (ξ, η) имеет распр. задаваемое плотностью $p(x,y) = p_{\eta,\xi}(x,y)$ тогда $p(x,y) = p(x/y)p(y) = p(y/x)p(x)$

Ф-ция $p(x/y) := \frac{p(x,y)}{p(y)}$, наз. усл. вер-ностью

Формула полной вер-ности $p(x) = \int_y p(x/y)p(y)dy$ или $p(x) = \sum_{y \in Y} p(x/y)p(y)$

Теорема Байеса $p(x/y) \to P(x/y) = \frac{p(y/x)P(x)}{p(y)} = \frac{p(y/x)P(x)}{\int_x p(y/x)p(x)dx}$

Предположения: Пусть известна совм. плотность распред. $p(x, y)$ на $X \times Y$, где $X = \mathbb{R}^n$, $Y = \{0, 1\}$, дай ф-ции потерь $L(a(x), y)$

Опр. Средняя бел. потеря для алгоритма $a(x)$:

$$R(a) = \int\int L(a(x), y)\, dP(x, y) = \int\int L(a(x), y)\, P(x, y)\, dx\, dy$$

Хотим решать задачу бинарной классификации, но наб. y x опр. описан y, Y.

Задача Найти такой алгоритм $a^*(x)$, где $a^*(x) = \arg\min_a R(a)$

Будем искать модель a^* оптимальной, R^* - значение оптимального среднего риска.

Билет 9

Лин. Б.К. Кли и квадр. дискр.

Байесовский классификатор. Теорема об оптимальности байесовского классификатора

Опр. Распределение принадл. экспоненциальному семейству распределений, если ин-кость распр. можно быть записана в следующем виде:

$$p(x|\theta) = h(x) g(\theta) e^{\xi(\theta)^T T(x)}$$

Пример: нормальное, энергии., биномиальное, Г-распр.

Предп. $T(x) = x$, $p(x|y) = h(x)\, g_y(\theta_y)\, e^{\xi_y(\theta_y) x}$

Теорема о лин. байесовском классификаторе: Если для бин. класс. плотности распр. имеют вид: $p(x|y)$, $h(x)\, g_y(\theta_y)\, e^{\xi_y(\theta_y) x}$, и среди признаков есть константа, то верно: 1) разделяющая поверхность $(w, x) = \ln \frac{\lambda_+}{\lambda_-}$ 2) $p(y|x) = \sigma(\langle w, x \rangle y)$, где $\sigma(z) = \frac{1}{1+e^{-z}}$ - лог. ф-ция.

Д-во: $\frac{p(y=+1|x)}{p(y=-1|x)} = \frac{p(x|y=+1)\, p(y=+1)}{p(x|y=-1)\, p(y=-1)} = \frac{p(y=+1)\, h(x)\, g_+(\theta_+)\, e^{\xi_+(\theta_+) x}}{p(y=-1)\, h(x)\, g_-(\theta_-)\, e^{\xi_-(\theta_-) x}} =$

$$= \frac{p(y=+1) \cdot g_+(\theta_+)}{p(y=-1) \cdot g_-(\theta_-)} e^{(\xi_+(\theta_+) - \xi_-(\theta_-)) x} = e^{\langle w, x \rangle}$$

Сделав предп. о const, выразим $\ln$ от e и получим $\langle w, x \rangle$

Из полученного вер. и того, что $p(y=+1|x) + p(y=-1|x) = 1 \Rightarrow$

$\Rightarrow p(y|x) = \sigma(\langle w, x \rangle, y)$, где $\sigma(z) = \frac{1}{1+e^{-z}}$ -

Для бин. класс. разделяющей поверхности оптимального байесовского классификатора имеем вид $\frac{p(y=+1|x)}{p(y=-1|x)} - \frac{\lambda_-}{\lambda_+} = e^{\langle w, x \rangle} - \frac{\lambda_-}{\lambda_+} = 0$

Классификатор Quadratic Discriminant Analysis QDA

- предп., что выбр. из каждого класса норм. распр.
- предп., что каждый класс имеет свою ковар. матр.

Классификатор Linear Discriminant Analysis LDA

- -//-
- -//- одинаковую
- -//- ковар. матр.

QDA $a(x) = \frac{1}{2} x^T A x + \langle w, x \rangle - b = 0$, где $A = \Sigma_0^{-1} - \Sigma_1^{-1}$

M_y - вектор матож. в классе y $w = M_1^T \Sigma_1^{-1} - M_0^T \Sigma_0^{-1}$

Σ_y - ковар. матр. реализ. x $\quad b = \ln \frac{P_1}{P_0} + \frac{1}{2} \ln \frac{\det \Sigma_0}{\det \Sigma_1} - \mu_1^T \Sigma_1^{-1} \mu_1 + \mu_0^T \Sigma_0^{-1} \mu_0$ (см. 5)

LDA: $a(x) = (w, x) - b = 0$, где $w = (\mu_1 - \mu_0)^T \Sigma^{-1}$

$b = \ln \frac{P_1}{P_0} - \frac{1}{2}(\mu_1 - \mu_0)^T \Sigma^{-1} (\mu_0 + \mu_1)$

Билет 9

Байесовский классификатор. Теорема об оптимальности байесовского классификатора

Виды ф-ций потерь для задачи бин. классификации:

1) cross-entropy $L = P(y=1) \log(a(x)=1) + P(y=0) \log(a(x)=$ logloss

2) индикатор $\begin{cases} 1, & a(x) = y \\ 0, & a(x) \neq y \end{cases}$

Рассм. 2) ф-цию потерь $L(a(x), y) = [a(x) \neq y]$

Средний риск $R(a) = \int\int L(a(x), y) p(x, y)\, dx\, dy = \int_X \sum_y [a(x) \neq y] P(x|y) P(y)\, dx$

$= \int_X \sum_y (1 - [a(x) = y]) P(x|y) P(y)\, dx = \int_X \sum_y P(x|y) P(y)\, dx - \int_X \sum_y [a(x) = y] P(x|y) P(y)\, dx$

Тогда следует $R(a) = \arg\max_a \int_X \sum_y [a(x) = y] P(x|y) P(y)\, dx = \arg\max_y P(x|y) P(y)$, $a^*(x)$

Пусть ф-ция потерь $L(a(x), y) = \lambda_y [a(x) \neq y]$, где $\lambda_y \geq 0$

Теорема Минимум средних потерь при ф-ции потерь $L(a(x), y)$ достигается байесовским классификатором

$a(x) = \arg\max_y \lambda_y (P(y|x)) = \arg\max_y \lambda_y P(y) P(x|y)$

Следствие: Оптимальное правило классификации при одинаковых штрафах за ошибку максимум. априорн. вер-сть класса

Билет 6

Наивный байесовский классификатор

Предп. Все признаки явл. неав. с.в. $p(x|y) = \prod_i p(x_i|y)$

Наивный байесовский классификатор: $a(x) = \arg\max_{y \in Y} P(y) \prod_i p(x_i|y)$

Если предполож., что $p_i(x_i|y) = \frac{1}{\sqrt{2\pi}\sigma_y} e^{\frac{-(x_i - \mu_y)^2}{2\sigma_y^2}}$, тогда классификатор будет наз. наивным байесовским гауссовским классиф.

Параметры байесовского класс. $(P(y), \mu, \sigma)$ настраиваемого по обуч. вы-ку.

Замечания: работает на малых выборках (из обр. матриц)

Недостатки: . расп. иногда не нрв.
. обуч. есть только обуч. выборка, т.е. реальн. расп.

Основные подходы:

. метод. точность расп. по входн. данным.
. сделать предп. о наз-ии семейства (ф. расп.) и по данным настроить пар-ры
. исп. методы, метода вычисл. плотность

Билет 10

Интервал. оценка вероятности плотности распр. Оценка Парзена-Розенблатта (одном. случай)

Задача: По выборке $X^l = (x_i)_{i=1}^l$ оценить пл-ость вероятности $p^T(x)$ (без вед. нер. плотность)

Дискр. случай: Признак x_i может принимать кон. число значений. Гистограммная оценка $\hat{p}(x) = \frac{1}{l} \sum_{i=1}^{l} [x = x_i]$

Непр. случай (одном.) По опр. пл-ти $p(x) = \lim_{h \to 0} \frac{1}{2h} P(x-h, x+h)$

Выберем x_i, $h \in R$ и положим h: $\hat{p}_h(x) = \frac{1}{2h} \frac{1}{l} \sum_{i=1}^{l} [|x - x_i| < h] = \frac{1}{2h} \frac{1}{l} \sum_{i=1}^{l} [\dots]$

$(\frac{x - x_i}{h} < 1)$. Эмпирическая оценка $\hat{p}_h(x)$ она h дел. пл-ть вер. доли объектов выборки

Оценка Парзена - Розенблатта

Эмпирическая оценка плотности по ону ширине h: $\hat{p}_h(x) = \frac{1}{l h} \sum_{i=1}^{l} \frac{1}{2}$ $\left[\frac{|x - x_i|}{h} < 1\right]$. Обобщ.: $\hat{p}_h(x) = \frac{1}{l h} \sum_{i=1}^{l} K\left(\frac{x - x_i}{h}\right)$

где $K(z)$ - ядро, удовл. требованиям: четная ф-ция, нормир. $\int_{-\infty}^{\infty} K(z)\,dz = 1$

Епанечн. ф-ция $z > 0$, неотр. ф-ция

Теорема. ($x_i \in \mathbb{R}$)

Пусть выполнено: x^l - простая выборка из распред. $p(x)$;

ядро $K(z)$ непр. и огр. $\int_{-\infty}^{\infty} K^2(z)\,dz < \infty$; посл. h_l: $\lim_{l \to \infty} h_l = 0$ и $\lim_{l \to \infty} l h_l = \infty$

Тогда 1) $\hat{p}_{h_l}(x) \xrightarrow{l \to \infty} p(x)$ для $x \in X$

2) скорость сход. имеет порядок $O(l^{-2/5})$

Билет 11

Постановка задачи регрессии. Аналитическое решение задачи линейной регрессии. Связь с методом наименьших квадратов

Постановка задачи:

Пусть X - объекты из $\mathbb{R}^n$, Y - ответы из $\mathbb{R}$

$x^l = (x_i, y_i)_{i=1}^l$ - обучающая выборка, $y_i = y(x_i)$, где $y: X \to Y$ - неизвестная

ищут. зависимость. $a(x) = f(x, w)$ - модель, $w \in \mathbb{R}^p$ - вектор парам. модели

Метод наименьших квадратов (в общ. виде)

$Q(w, x^l) = \sum_{i=1}^{l} (\alpha_i (f(x_i, w) - y_i)^2) \to \min_w$, где α_i - вес, степень важности x_i

$Q(w, x^l)$ - эмп. сумма квадратов

Предположим, что $a(x) = f(x, w) = w_0 + w_1 x^1 + \dots + w_n x^n$, где $w = (w_0, \dots, w_n)^T \in \mathbb{R}^{n+1}$ параметры модели

Другая форма: $a(x) = \underbrace{w^T \cdot x}_{\text{скал. п.}}$, где $x = (1, x^1, \dots x^n)^T \in \mathbb{R}^{n+1}$

Метод наим. квадратов: $L(w, X_{train}) = MSE(w, X_{train}) = \frac{1}{l} \sum_i (w^T x_i - y_i)^2$ → ф-ция потерь

Найти $\hat{w} = \arg\min_w (L(w, X_{train}))$

Теорема: Решением задачи $\arg\min_w \left(\sum (w^T x_i - y_i)^2\right)$ обл. $\hat{w} = (X^T X)^{-1} X^T y$, где $X = \{x_i^j\}$ - номер ранга, $y = (y_1, \dots y_\ell)$

Д-во: $\|Xw - y\|^2 \to \min_w$

$\frac{\partial}{\partial w}\|Xw - y\|^2 = 0 \Rightarrow \frac{\partial}{\partial w}\|Xw - y\|^2 = \frac{\partial}{\partial w}((Xw-y)^T(Xw-y)) = \frac{\partial}{\partial w}((Xw)^T Xw -$

$- (Xw)^T y - y^T Xw + y^T y) = \frac{\partial}{\partial w}(w^T X^T Xw - w^T X^T y - y^T Xw + y^T y) = \frac{\partial}{\partial w} w^T (X^T X) w -$

$- 2\frac{\partial}{\partial w}(X^T y)^T w = 0$

Опр Пусть $w = (w_1, \dots w_n)$, $z = z(w_1, \dots w_n)$. Тогда $\frac{\partial z}{\partial w} := \left(\frac{\partial z}{\partial w_1}, \dots \frac{\partial z}{\partial w_n}\right)^T$

Лемма: $\frac{\partial}{\partial x} x^T a = a$

Лемма $\frac{\partial}{\partial x} x^T A x = (A + A^T)x$

$\frac{\partial}{\partial w} w^T (X^T X) w = 2\frac{\partial}{\partial w}(X^T y)^T w = 2X^T Xw - 2X^T y = 0 \Rightarrow w = (X^T X)^T X^T y$

Билет 12

Регрессия и метод наименьших квадратов (МНК). Теорема об эквивалентности решений МНК и МПП

Модель данных с центрированным гауссовским шумом:

$y(x_i) = f(x_i; w) + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma_i)$, $i = 1, \dots \ell$

Метод максимума правдоподобия:

$L(\varepsilon_1, \dots \varepsilon_\ell | w) = \prod_{i=1}^{\ell} \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{\varepsilon_i^2}{2\sigma_i^2}} \to \max_w$

$-\log L(\varepsilon_1, \dots \varepsilon_\ell | w) = \text{const}\, w + \frac{1}{2} \sum_{i=1}^{\ell} \frac{1}{\sigma_i^2} F(x_i, w) - y_i)^2 \to \min_w$ - метод наим. кв.

Теорема: При предпол. выше (про гаус. шум) решения МНК и ММП совп., причём веса обратно пропорц. дисперсии шума $w_i = \sigma_i^2$

Билет 13

Теорема о минимальности величины среднего потерь для регрессии с ф-цией ошибки MSE.

Пусть известно совм. распр. $p(x, y)$ на $X \times Y$ (x - приц., y - ответы)

Пусть задана ф-ция потерь $L(a(x), y)$

Опр. Средняя величина потерь для алгоритма $a(x)$

$R(a) = \int\int L(a(x), y)\, dP(x, y) = \int\int L(a(x), y) p(x, y) dx dy$

Задача. Найти $a^*(x)$: $a^*(x) = \arg\min_a R(a)$, где a^* - оптим. модель, R - значение оптимального среднего риска

Теорема: Если $L(a(x), y) = (a(x) - y)^2$, то сред. средних потерь минимальна при $a^* = E(y|x)$ - усл. матожид.

Д-во. $R(a) = \int\int L(a(x), y) p(x, y)\, dy\, dx = \int\int (a(x) - y)^2 p(x, y)\, dy dx =$

$= \int\int (a(x) - y)^2 p(y|x)\, dy\, p(x) dx = \int E((y - a(x))^2 | x) p(x) dx$

Лемма $E((y - a(x))^2 | x) = E((y - E(y|x))^2 | x) + E((a(x) - E(y|x))^2 | x)$

$R(a) = \int E((y - a(x))^2 | x) p(x) dx = \int E((y - E(y|x))^2 | x) p(x) dx +$

$+ \int E((a(x) - E(y|x))^2 | x) p(x) dx \geq \int E((y - E(y|x))^2 | x) p(x) dx$

Билет 14

Решение задачи гребневой регрессии. Вероятностный смысл регуляризации. L_1, L_2, Elastic-Net регуляризации (без теорем)

L_2-регуляризация: $L(w, X_{train}) = MSE(w, X_{train}) + \frac{\alpha}{2}\sum_{i=0}^{n} w_i^2 = \frac{1}{\ell}\sum_i (w^T x^i - y_i)^2 + \frac{\alpha}{2}\sum_{i=1}^{n} w_i^2$ — ф-ция потерь

Задача: найти $\hat{w} = \underset{w}{\arg\min}(L(w, X_{train}))$

Теорема: Решением задачи $\underset{w}{\arg\min}(\sum_i (w^T x^i - y_i)^2 + \alpha \sum_{i=0}^{n} w_i^2)$ явл. $\hat{w} = (X^T X + \alpha I_{n+1})^{-1} X^T y$, где $x_{ij} = x_i^j$, $y = (y_1 \dots y_\ell)$, I_{n+1} — ед. матр.

Лемма $\frac{\partial}{\partial x} x^T x = 2x$

$\|Xw - y\|^2 + \alpha\|w\|^2 \to \min_w$

Необх. усл. минимума: $\frac{\partial}{\partial w}((Xw-y)^T(Xw-y) + \alpha w^T w) = 2X^T X w - 2X^T y + 2\alpha w = 0$

Свойства: • регуляризация не дает пар-ам модели быть слишком большими

• Имеет хорошую обобщ. способность, как правило, рег. обесп. большую обобщ. способность

• Более устойчива к выбросам

• Появился параметр, который можно настроить при помощи кросс-валидации

Вер. смысл α: $\alpha = \frac{1}{T^2}$, T — среднекв. отклон. априорного распр. на w

L_1-регуляризация: $L(w, X_{train}) = MSE(w, X_{train}) + \alpha\sum_{i=0}^{n}|w_i| = \sum_i (w^T x^i - y_i) + \alpha\sum_{i=0}^{n}|w_i|$ — ф-ция потерь

Задача: найти $\hat{w} = \underset{w}{\arg\min}(L(w, X_{train}))$, нет аналит. реш.

С-ва: Обеспечивает отбор признаков

Смысл α: α обр. пропорц. среднеквадр. отклон. апр. распр. ω. Лапласа $p(w) = \frac{1}{T} e^{-\frac{\|w\|}{2T}}$

В данном случае это распр.

Elastic Net: $L(w, X_{train}) = MSE(w, X_{train}) + 2\alpha\sum_{i=0}^{n}|w_i| + (1-2)\frac{\alpha}{2}\sum_{i=0}^{n} w_i^2 = \sum_i (w^T x^i - y_i)^2 + 2\alpha\sum|w_i| + (1-2)\frac{\alpha}{2}\sum w_i^2$ — ф-ция потерь

Задача: Найти $\hat{w} = \underset{w}{\arg\min}(L(w, X_{train}))$

С-ва: Нет аналитического решения, совм. св-ва Ridge и Lasso рег.

Билет 15

Непараметрическая регрессия: Формула Надарая-Ватсона (б/д)

Постановка зад. регр. и допущения

• X — объекты (часто $\mathbb{R}^n$), Y — ответы (часто $\mathbb{R}$), на X р-метрика $y: X \to Y$ неизв. завис.

• $X^\ell = (x_i, y_i)_{i=1}^{\ell}$ — обуч. выборка, $y_i = y(x_i)$,

• $a(x) = f(x, w)$ — пар-ая модель зав-сти из $\mathbb{R}^p$ — вектор-параметров

• Метод наим. квадр: $Q(w, X^\ell) = \sum(\alpha_i (f(x_i, w) - y_i)^2) \to \min_w$ где α_i — вес, степень важности объекта;

Недостаток: надо иметь хорошую парам. модель $f(x, w)$

Будем прибл. модель const $f(x, w) = w$ в окр-ти $x \in X$

(стр. 9)

$Q(w, X^l) = \sum (\alpha_i(x)(w - y_i)^2) \to \min_w$, где $\alpha_i(x) = K\left(\frac{\rho(x, x_i)}{h}\right)$ - веса объектов x_i относ. x, $K(z)$ - ядро, невозр., огр., гладкое, h - ширина окна сглаживания

Формула Н-В: $a_h(x, X^l) = \frac{\sum_{i=1}^{l} y_i \alpha_i(x)}{\sum_{i=1}^{l} \alpha_i(x)} = \frac{\sum_{i=1}^{l} y_i K\left(\frac{\rho(x, x_i)}{h}\right)}{\sum_{i=1}^{l} K\left(\frac{\rho(x, x_i)}{h}\right)}$

Теорема. Пусть выполнены усл.

1) Выборка $X^l = (x_i, y_i)_{i=1}^{l}$ прост. из распр. $p(x, y)$
2) Ядро $K(z)$ огр. и $\int_0^\infty K(z)\,dz < \infty$, $\lim_{z \to \infty} zK(z) = 0$
3) Зависимость $E(y|x)$ не имеет верт. асимптот: $E(y^2|x) = \int_Y y^2 p(y|x)\,dy < \infty$ при $\forall x \in X$
4) Посл. h_l убывает, но не слишком: $\lim_{l \to \infty} h_l = 0$, $\lim_{l \to \infty} l h_l = \infty$

Тогда имеет место сход. по вер-ти: $a_{h_l}(x, X^l) \xrightarrow{P} E(y|x)$ в любых $x \in X$, где $E(y|x)$, $p(x)$, $D(y|x)$ непр. и $p(x) > 0$

Виды ядер: $\Pi(z) = [|z| \le 1]$ - прямоуг. $E(z) = (1 - z^2)[|z| \le 1]$ - квадратичное

$Q(z) = (1 - z^2)^2 [|z| \le 1]$ - квартическое

$T(z) = (1 - |z|)[|z| \le 1]$ - треугольное

$G(z) = e^{-2z^2}$ - гауссовское

$G(z)$ - гладкая аппроксимация, ширина окна существенно влияет на точность аппрокс.

$T(z)$ - кус. лин. аппр., не огр.

$\Pi(z)$ - кус. пост. аппр., в случае отсутствия т. выборки выбор ядра слабо влияет на точность.