Задачи к лекции и семинару “Регуляризация”

1

Пусть дана выборка точек на прямой $\{x_i\}$. Максимизируйте правдоподобие (или его логарифм) в гауссовой вероятностной модели:

$$\prod_i p(x_i) \to \max_{\mu,\sigma} \qquad p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

2

Количество срабатываний счетчика Гейгера за минуту n подчиняется распределению Пуассона:

$$P_\lambda(n) = \frac{\lambda^n}{n!} e^{-\lambda}.$$

1. В ходе эксперимента счетчик Гейгера сработал за минуту m раз. С помощью теоремы Байеса определите апостериорное распределение на λ. *Указание*: априорную плотность вероятности λ можно считать постоянной (так как мы изначально ничего не знаем про λ)[1].
2. Эксперимент повторили еще раз, в этот раз счетчик Гейгера сработал за минуту m' раз. Как обновилось апостериорное распределение на λ?

3

Ультрачувствительный тест от короновируса ошибается в 1% случаев (как в одну, так и в другую сторону). В данный момент в популяции доля заболевших 10^{-5}. Петя получил положительный тест на коронавирус. С какой вероятностью он действительно болеет коронавирусом?

4

Пусть имеется априорное распределение на вектор $\boldsymbol{x}$, задаваемое матрицей A:

$$p_0(\boldsymbol{x}) = \frac{1}{Z} e^{-\frac{\boldsymbol{x}^T A \boldsymbol{x}}{2}}.$$

Было произведено измерение величин $\boldsymbol{x}$, которое дало значение $\boldsymbol{x}_1$. Найдете апостериорное распределение на $\boldsymbol{x}$.

5

На семинаре обсуждалось решение задачи регрессии с $L1$-регуляризацией с помощью метода градиентного спуска. С помощью K-Fold кроссвалидации[2] ($K = 3$) осуществите для этого метода подбор параметров: коэффицент перед регуляризатором и параметр градиентного спуска (learning rate). В качестве данных возьмите значения какой-нибудь неполиномиальной функции на равномерной или случайной сетке (на выбор семинариста) с добавленным гауссовым шумом. Насколько стабильно по отношению к запуску работает градиентный спуск?

---

[1] Такая плотность вероятности не будет нормируема. Чтобы сделать рассуждение более строгим, можно ввести обрезку на очень больших λ (так как это нереалистичные значения). Другими словами, можно считать, что априорная плотность вероятности $p_0(\lambda)$ — это какая-то очень медленно меняющаяся функция и как-то убывающаяя на бесконечности. Тогда в числителе и знаменателе формулы Байеса она будет домножаться на гораздо более быструю функцию и поэтому можно заменить $p_0(\lambda) \to p_0(0)$. Константа $p_0(0)$ должна сократиться в ходе вычислений.

[2] Способ кроссвалидации, при котором данные разбиваются на K частей и совершается K запусков обучения: по очереди каждая из частей объявлется тестовой, а объединение оставшихся $K - 1$ частей используется для обучения. Такой способ обсуждался на первой лекции.

6

Для стандартного набора данных для задачи регрессии (см. например `load_diabetes` из `sklearn.datasets`) продемонстрируйте, как веса обращаются в ноль по мере увеличения коэффицинета $L1$-регуляризации. Разрешается использовать библиотечную реализацию регрессии.

7

Покажите, что задача минимизации квадратичной функции потерь с дополнительным ограничением (лассо Тибширани):

$$\mathcal{L} = \|Xw - y\|^2 \to \min_w, \qquad \sum_\alpha |w_\alpha| < C$$

эквивалентна L1-регуляризации. *Указание*: можно воспользоваться условиями Каруша — Куна — Таккера (обобщение метода Лагранжа). (link).

8*. Bias-Variance decomposition

Воспользуемся вероятностной моделью данных, в которой предполагается, что каждый элемент выборки независимо от других поступает из распределения $p(x, y)$. Тогда вероятность получить какой-то конкретный набор данных $(X_l, y_l) = (x_1, \ldots x_l; y_1, \ldots, y_l)$ в обучающей выборке равна $p(X_l, y_l) = \prod_{i=1}^l p(x_i, y_i)$. В дальнейшем будем обозначать как (x, y) элемент тестовой выборки, который не входит в (X_l, y_l).

В выбранной модели $\check{y} = g_\theta(x)$ параметры θ определяются с помощью фиттирования по обучающей выборке: $\theta = \theta(X_l, y_l)$, поэтому $\check{y}$ зависит от x, X_l и y_l. Тогда формальное выражение для функции потерь (соответствующее пределу бесконечной большой тестовой выборки) можно записать как

$$L = \mathbb{E}_{X_l, y_l}\left[\mathbb{E}_{x,y}\left(y - \check{y}\right)^2\right].$$

В этом выражении квадратичная функция потерь усредняется по элементу тестовой выборки (x, y) и по обучающей выборке (X_l, y_l).

Покажите, что справедливо разложение этой величины на шум, смещение и разброс:

$$L = \underbrace{\mathbb{E}_{x,y}\left(y - \mathbb{E}(y|x)\right)^2}_{\text{noise}} + \underbrace{\mathbb{E}_{x,y}\left(\mathbb{E}_{X_l, y_l}(\check{y}) - \mathbb{E}(y|x)\right)^2}_{\text{bias}} + \underbrace{\mathbb{E}_{x,y}\left[\mathbb{E}_{X_l, y_l}\left(\check{y} - \mathbb{E}_{X_l, y_l}\ \check{y}\right)^2\right]}_{\text{variance}}$$

Указание: сначала покажите, что

$$\mathbb{E}_{x,y}\left(y - \check{y}\right)^2 = \mathbb{E}_{x,y}\left(y - \mathbb{E}(y|x)\right)^2 + \mathbb{E}_{x,y}\left(\mathbb{E}(y|x) - \check{y}\right)^2$$