# Домашнее задание по материалу 2-го семинара. ММП, весна 2013 2 марта

1. Бэггинг, как было рассказано на семинаре, использует в своей работе ряд выборок $\tilde{X}_m^\ell$ длиной $\ell$, сгенерированных из обучающей выборки $X^\ell$. Каждая очередная выборка $\tilde{X}_m^\ell$ независимо генерируется по следующей схеме: из обучающей выборки $\ell$ раз независимо, равномерно с возвращениями вытягиваются объекты. Такие выборки называются *бутстреп выборками (bootstrap samples)* — отсюда и название метода bagging: bootstrap aggregation.

   Вычислите вероятность того, что объект $x_i \in X^\ell$ попадет в выборку $\tilde{X}_m^\ell$. Чему она равна при $\ell \to \infty$?

2. Bootstrap выборки часто используются для изучения свойств статистических оценок. Под статистическими оценками подразумеваются любые функции выборок (наблюдений случайных величин), которые используются для оценивания тех или иных параметров распределений случайных величин. Например, выборочная дисперсия, посчитанная для простой выборки $X^n = (x_1, \ldots, x_n)$, — это статистика, оценивающая дисперсию распределения, сгенерировавшего эти наблюдения.

   Рассмотрим произвольную статистику $S\colon X^n \to \mathbb{R}$ (например, $S(X^n)$ может быть выборочным средним). Предположим, что у нас есть $M$ бутстреп выборок $\tilde{X}_m^n$, $m = 1, \ldots, M$, полученных из выборки $X^n$. Используя их, мы можем оценивать любые свойства статистики $S$ — например, дисперсию статистики:

$$\widehat{\mathrm{Var}}[S(X^n)] = \frac{1}{M-1} \sum_{m=1}^{M} \left( S(\tilde{X}_m^n) - \frac{1}{M} \sum_{i=1}^{M} S(\tilde{X}_i^n) \right)^2 .$$

   Это выражение можно в свою очередь рассматривать, как приближение (оценку *Монте-Карло*) для дисперсии $\mathrm{Var}_{\tilde{X}^n}[S(\tilde{X}^n)]$, где случайная выборка $\tilde{X}$ распределена описанным выше способом (каждый ее элемент вытягивается независимо, равномерно и с возвращениями из выборки $X^n$).

   Описанный метод оценивания называется *бутстреппинг*. Он очень хорошо изучен в теории вероятностей и математической статистике, широко применяется и у него есть много полезных свойств.

   Попробуем использовать бутстреппинг для изучения $\mathbb{E}_{x,y}[h^\ell(x) \neq y]$ — среднего риска классификатора $h^\ell$, настраевомого по обучающей выборке $X^\ell$ некоторым

фиксированным методом обучения $\mu\colon X^\ell \to \mathcal{H}$. Как мы вскоре убедимся, это не лучшая идея.

Рассмотрим следующий искусственный пример. Дана задача классификации на два класса $\mathbb{Y} = \{-1, +1\}$ и обучающая выборка $X^\ell$. Предположим, что ответы $Y$ и объекты $X$ независимы и $P(Y = +1) = \frac{1}{2}$. Мы будем использовать алгоритм одного ближайшего соседа $h_{1\text{NN}}(x)$ для решения этой задачи. Попробуем оценить его средние потери $\mathbb{E}_{x,y}[h_{1\text{NN}}(x) \neq y]$ с помощью бутстрепинга, то есть оценим их с помощью

$$\widehat{Err}(X^\ell, \mu) = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{\ell}\sum_{i=1}^{\ell}[y_i \neq h^m_{1\text{NN}}(x_i)]\right),$$

где $h^m_{1\text{NN}}$ — классификатор, настроенный по $m$-ой бутстреп выборке $\tilde{X}^\ell_m$. Это один из возможных *критериев выбора* метода обучения $\mu$.

а) Чему равно истинное значение $\mathbb{E}_{x,y}[h_{1\text{NN}}(x) \neq y]$?

б) Какое значение будет иметь матожидание $\mathbb{E}_{X^\ell}\widehat{Err}(X^\ell, \mu)$? При ответе на этот вопрос вам пригодится результат задачи 1.

в) Подумайте над тем, в чем заключается отличие критерия скользящего контроля от приведенного выше бутстреп критерия? Почему бустреппинг работает хуже?

3. Предположим, что в задаче регрессии $\mathbb{Y} = \mathbb{R}$ с обучающей выборкой $X^\ell$ распределение $P(X, Y)$ таково, что $Y = f(X)+\varepsilon$, где случайная величина $\varepsilon$ независима от $X$, $\mathbb{E}[\varepsilon] = 0$ и $\text{Var}[\varepsilon] = \sigma^2$. Также для простоты предположим, что объекты $x_1, \ldots, x_\ell$ обучающей выборки зафиксированы и случайность обучающей выборки обусловлена лишь случайностью ответов на точках обучающей выборки. Для решения этой задачи мы будем использовать алгоритм $K$-ближайших соседей, то есть

$$h(x) = \frac{1}{K}\sum_{k=1}^{K} y_{(k)},$$

где $y_{(k)}$ — ответ на $k$-ом соседе $x$ из обучающей выборки.

а) Вычислите рассмотренное на семинаре разложение величины

$$\mathcal{L}(\mu) = \mathbb{E}_{X^\ell}\left[\mathbb{E}_{x,y}\left[\left(y - h^\ell(x)\right)^2\right]\right], \text{ где } h^\ell = \mu\left(X^\ell\right)$$

на шум, отклонение и дисперсию

$$\mathcal{L}(\mu) = \underbrace{\mathbb{E}_{x,y}\left[(y - \mathbb{E}[y|x])^2\right]}_{\text{noise}} + \underbrace{\mathbb{E}_x\left[\left(\mathbb{E}[y|x] - \mathbb{E}_{X^\ell}\left[h^\ell(x)\right]\right)^2\right]}_{(\text{bias})^2} +$$
$$+ \underbrace{\mathbb{E}_x\left[\mathbb{E}_{X^\ell}\left[\left(h^\ell(x) - \mathbb{E}_{X^\ell}\left[h^\ell(x)\right]\right)^2\right]\right]}_{\text{variance}}$$

для алгоритма $K$ ближайших соседей.

б) Пользуясь полученным разложением исследуйте поведение метода ближайших соседей при изменении K. Убедитесь, что K является параметром *сложности* модели: с уменьшением K сложность возрастает, а значит, дисперсия увеличивается и отклонение падает. И наоборот.