Школа анализа данных
Машинное обучение, часть 2
Домашнее задание №2

Решите предложенные задачи. Решения необходимо оформить в виде PDF документа. Каждая задача должна быть подробно обоснована, задачи без обоснования не засчитываются. Решения пишутся в свободной форме, однако так, чтобы проверяющие смогли разобраться. Если проверяющие не смогут разобраться в решении какой-нибудь задачи, то она автоматически не засчитывается. Дедлайн очников 30 октября 2017 09:00MSK, дедлайн заочников и филиалов +2 суток.

Задача 1 (2 балла) Bias-variance decomposition.

Пусть задана выборка $X^\ell = \{x_1, \dots, x_\ell\} \subset \mathbb{X}$ с вещественными ответами $y^\ell = \{y_1, \dots, y_\ell\}$ из $\mathbb{Y} = \mathbb{R}$. Будем считать, что на пространстве всех объектов и ответов $\mathbb{X} \times \mathbb{Y}$ существует распределение $p(x, y)$, из которого была сгенерирована заданная выборка X^ℓ и ответы y^ℓ на ней. Пусть также задан некоторый метод обучения μ. Обученный этим методом на обучающей выборке алгоритм будем обозначать через $\mu(X^\ell, y^\ell)$. Ответ обученного алгоритма на объекте x будем обозначать через $\mu(X^\ell, y^\ell)(x)$. Далее для сокращения записей вместо $\mu(X^\ell, y^\ell)(x)$ будем писать просто $\mu(x)$.

Выведите разложение на смещение и разброс (bias-variance decomposition) для матожидания среднеквадратичной ошибки метода обучения $\mu(X^l)$, выполнив описанные ниже шаги.

$$
\begin{aligned}
L(\mu) &= \mathbb{E}_{X^\ell, y^\ell}\Big[\mathbb{E}_{x,y}\Big[\big(y - \mu(x)\big)^2\Big]\Big] = \\
&= \underbrace{\mathbb{E}_{x,y}\Big[\big(y - \mathbb{E}[y \,|\, x]\big)^2\Big]}_{\text{шум}} + \\
&+ \underbrace{\mathbb{E}_{x,y}\Big[\big(\mathbb{E}_{X^\ell, y^\ell}\big[\mu(x)\big] - \mathbb{E}[y \,|\, x]\big)^2\Big]}_{\text{смещение}} + \underbrace{\mathbb{E}_{x,y}\Big[\mathbb{E}_{X^\ell, y^\ell}\Big[\big(\mu(x) - \mathbb{E}_{X^\ell, y^\ell}\big[\mu(x)\big]\big)^2\Big]\Big]}_{\text{разброс}}.
\end{aligned}
$$

1. Докажите, что имеет место следующее разложение среднеквадратичного риска на фиксированной выборке X^ℓ:

$$\mathbb{E}_{x,y}\Big[\big(y - \mu(x)\big)^2\Big] = \mathbb{E}_{x,y}\Big[\big(y - \mathbb{E}[y \,|\, x]\big)^2\Big] + \mathbb{E}_{x,y}\Big[\big(\mathbb{E}[y \,|\, x] - \mu(x)\big)^2\Big].$$

   Подсказка. $\mathbb{E}[y \,|\, x]$ не зависит от y, только от x.

2. Подставьте это разложение в $L(\mu)$.

3. Получите шумовую компоненту разложения, найдя слагаемое, не зависящее от X^ℓ.

4. Разложите второе слагаемое на три, добавив и вычтя $\mathbb{E}_{X^\ell, y^\ell}[\mu(x)]$ внутри квадрата.

5. Покажите, что слагаемое

$$\mathbb{E}_{X^\ell, y^\ell}\Big[\big(\mathbb{E}[y \,|\, x] - \mathbb{E}_{X^\ell, y^\ell}\big[\mu(x)\big]\big)\big(\mathbb{E}_{X^\ell, y^\ell}\big[\mu(x)\big] - \mu(X^\ell)\big)\Big]$$

   равно нулю.

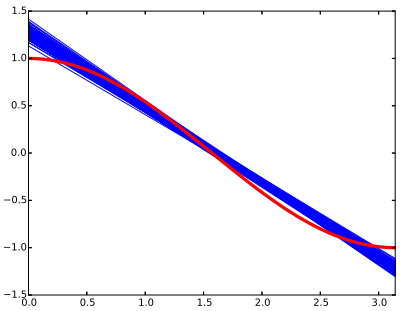

Рис. 1: Линейная регрессия

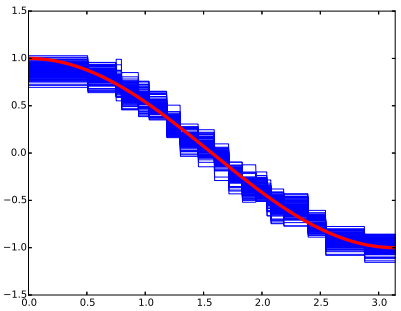

Рис. 2: kNN, $k = 2$

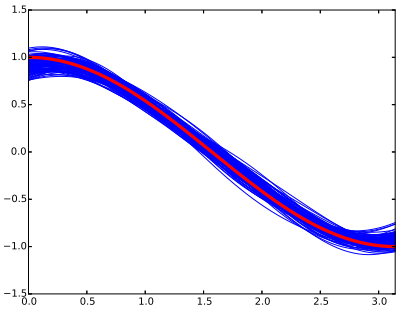

Рис. 3: SVM с RBF-ядром, $\gamma = 1$

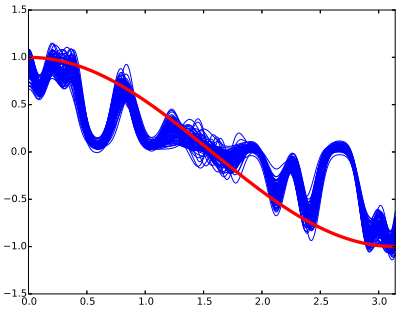

Рис. 4: SVM с RBF-ядром, $\gamma = 100$

Задача 2 (1.5 балла) Bias-variance decomposition.

Сгенерируем следующую выборку: объектами будут являться $\ell = 20$ фиксированных точек $X = \{x_1, \ldots, x_\ell\}$, равномерно распределенных на отрезке $[0, \pi]$, а ответами $y_i = \cos(x_i) + \mathcal{N}(0, 0.1)$. На рис. 1-4 для разных семейств алгоритмов изображена истинная кривая зависимости ответов от объектов $y = \cos(x)$, а также кривые для 100 алгоритмов, обученных на различных реализациях выборки. Охарактеризуйте смещение и разброс для каждого из четырех методов. Для более точной оценки смещения и разброса воспроизведите описанный эксперимент и посчитайте их численные значения.

Для описанной формулировки задачи некорректно говорить о смещении и разбросе на всем отрезке $[0, \pi]$. Объясните, почему. Предложите, как можно исправить условие, чтобы избежать этой некорректности. Перестройте графики к задаче для исправленного условия, и посчитайте смещение и разброс в новом эксперименте для каждого из четырех методов.

Задача 3 (2 балла) Ранжирование.

Имеются два ранжирующих правила $h_1(x)$ и $h_2(x)$. Дана выборка из одного запроса и N документов. Релевантности документов и ответы каждого из ранжирующих правил на каждом документе этой выборки известны. Предложите точный алгоритм построения линейной композиции

$$h(x) = \alpha h_1(x) + (1 - \alpha)h_2(x), \quad \alpha \in [0, 1]$$

двух ранкеров, имеющей оптимальное значение NDCG на этой выборке. Сложность алгоритма должна быть не выше $O(N^3)$.

Задача 4 (2.5 балла) Рекомендательные системы, матричные разложения.

Допустим, мы хотим обучить рекомендательную систему, имея историю взаимодействия n пользователей и m товаров. Рассмотрим матрицу оценок $R \in \mathbb{R}^{n \times m}$, в которой известны лишь некоторые элементы: r_{ui} – оценка пользователя $1 \leq u \leq n$ для товара $1 \leq i \leq m$. Дополнительно предположим, что доля пар (u, i) с известными оценками относительно всех пар равна α, причем эти пары равномерно распределены по матрице оценок. Рассмотрим способ предсказания неизвестных оценок на основе модели Latent Factor Model, которая использует малоранговую аппроксимацию ранга r матрицы R:

$$\tilde{r}_{ui} = p_u^T q_i, \quad p_u \in \mathbb{R}^r, \quad q_i \in \mathbb{R}^r.$$

Параметры модели можно подбирать в ходе решения задачи восстановления пропущенных значений матрицы оценок (Matrix Completion):

$$P, Q = \underset{P \in \mathbb{R}^{r \times n}, Q \in \mathbb{R}^{r \times m}}{\arg\min} \sum_{(u,i) \in \text{known}} (r_{ui} - p_u^T q_i)^2.$$

Рассмотрим один из алгоритмов решения этой оптимизационной задачи – Alternating Least Squares (ALS). Этот алгоритм поочередно фиксирует матрицы P или Q и подбирает оптимальное значение другой матрицы, точно решая задачу с помощью линейного метода наименьших квадратов. Например, при фиксированном P:

$$Q = \underset{Q \in \mathbb{R}^{r \times m}}{\arg\min} \sum_{(u,i) \in \text{known}} (r_{ui} - p_u^T q_i)^2.$$

Выпишите явные формулы для вычисления оптимального Q при фиксированном P, а также покажите, что вычислительная сложность одного такого шага равна $O\left(r^2 m(\alpha n + r)\right)$.

Задача 5 (1 балла) Рекомендательные системы, матричные разложения.

В факторизационных машинах предсказание для объекта $x \in \mathbb{R}^d$ делается по формуле

$$a(x) = w_0 + \sum_{i=1}^{d} w_j x_j + \sum_{i=1}^{d} \sum_{j=i+1}^{d} x_i x_j \langle \vec{v}_i, \vec{v}_j \rangle.$$

Вычисление предсказания по этой формуле требует $O(rd^2)$ операций, где r — размер векторов $\vec{v}_1, \ldots, \vec{v}_d$. Покажите, что это же предсказание может быть найдено за $O(rd)$ операций.

Задача 6 (1 балл) Рекомендательные системы.

Андрей и Денис решили воспользоваться машинным обучением, чтобы улучшить продажи в своем интернет-магазине. Для этого они разработали рекомендательную систему, предлагающую пользователю три товара в момент, когда он сформировал корзину и собрался делать заказ. Эта система делает рекомендации как на основе товаров в корзине, так и на основе истории пользователя. Например, если пользователь хочет купить фотоаппарат, то система предложит ему сумку и штатив, поскольку их часто покупают в дополнение к фотоаппаратам. Или же, если данный пользователь часто покупает книги, то система может предложить ему купить вместе с фотоаппаратом недавно вышедший детектив.

Рекомендательная система у Андрея и Дениса получилась очень сложная и ресурсоемкая, и для ее стабильной работы необходимо закупить несколько мощных серверов, а также платить зарплату системному администратору, который будет следить за работой этих серверов и самой рекомсистемы. Чтобы проверить, имеет ли смысл идти на такие траты, они хотят понять, способен ли разработанный алгоритм делать правильные рекомендации.

Денис предлагает взять историю покупок пользователей, и для каждого пользователя разбить ее на две части: первые 70% взять в обучающую выборку, а последние 30% — в контроль. После этого следует настроить рекомендательный алгоритм по обучающей выборке и проверить, насколько хорошо он предсказывает пользователям покупки из контрольной выборки.

Андрей же утверждает, что нужно провести AB-тест: разбить всех пользователей интернет-магазина на две группы и показывать рекомендации лишь в одной из групп. После этого предлагается подсчитать число купленных товаров в первой и второй группах. Если в группе, где показывались рекомендации, это число окажется больше, то рекомендательную систему следует признать полезной.

Какие преимущества и недостатки вы видите в подходах Андрея и Дениса? Кому из них следует отдать предпочтение?