Машинное обучение 1
Контрольная работа
Вариант 0

Задача 1 (2.5 балла). Ответьте на вопросы по логистической регрессии и SVM для задачи бинарной классификации:

1. Запишите модель логистической регрессии. Как в ней определить, к какому классу относится объект? Как определить вероятность принадлежности положительному классу?

2. Можно утверждать, что логистическая регрессия корректно оценивает вероятности принадлежности положительному классу. Запишите формальное определение для этого. Какую функцию потерь нужно подставлять в это определение, чтобы доказать, что у логистической регрессии всё хорошо с оценками вероятности?

3. Запишите условную задачу оптимизации для обучения SVM.

4. К сожалению, SVM плохо оценивает вероятности классов. Как можно исправить это? Опишите любой из двух способов, разобранных на занятиях.

5. Рассмотрим обычную логистическую регрессию без регуляризации и SVM. Допустим, мы применяем их к малой выборке, на которой число признаков примерно совпадает с числом объектов. У которого из двух методов будет выше смещение? У кого будет выше разброс?

Задача 2 (2.5 балла). Ответьте на вопросы по композициям:

1. Допустим, мы обучили градиентным бустингом композицию из 1000 деревьев. Композиция хорошо работает, но шпионы узнали, что наши конкуренты решают задачу на этих же данных с помощью того же градиентного бустинга, и им хватает 500 деревьев. Мы тоже хотим сократить размер композиции, и поэтому собираемся выбросить часть деревьев. Из лекций по МО-1 мы помним, что градиентный бустинг — это градиентный спуск в пространстве функций. Из тех же лекций мы помним, что первые шаги градиентного спуска происходят далеко от оптимума и дают плохие модели, а ближе к концу процесса мы уже получаем хорошие значения параметров. Значит, нужно выкинуть первые 500 деревьев из нашей композиции! Объясните, почему такой подход приведёт к тому, что конкурентам мы проиграем, а наши оценки за МО-1 аннулируют. Предложите, как более грамотно получить композицию из 500 деревьев, если нам разрешено заново обучить всё.

2. В одном из вариантов градиентного бустинга при выводе шагов используется разложение функции потерь в ряд Тейлора. Запишите, что именно раскладывается в ряд (с формулами!), и выведите итоговую задачу оптимизации для обучения очередного базового алгоритма (в общем случае, не для деревьев).

3. Для чего при обучении случайного леса семплируются признаки в каждой вершине? Обоснуйте с точки зрения смещение, разброса и шума.

Задача 3 (2.5 балла). Пусть у нас есть выборка $X = \{(x_i, y_i)\}_{i=i}^{\ell}$ для бинарной классификации, в которой доля объектов q класса $+1$ намного меньше доли объектов $(1-q)$ класса -1. Как обычно, мы решаем задачу

$$Q(X, a) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i, a(x_i)) \to \min_a$$

.

- Обозначим через $\bar{L}_y$ среднее функции потерь на объектах класса y. Выразите $Q(X, a)$ через $\bar{L}_{+1}$ и $\bar{L}_{-1}$ и проинтерепретируйте полученную формулу.
- Основываясь на предыдущем пункте, скажите, можно ли ожидать хорошее качество от такого классификатора? Почему?
- Предложите несколько способов (как минимум два) решения проблемы.
- Предположим, что при сборе данных произошла ошибка, и на самом деле при правильном сэмплировании объектов доля класса $+1$ была бы равна q', а класса -1, соответственно, — $(1-q')$. Предположим также, что среднее функции потерь по объектам одного класса — достаточно хорошая оценка и не сильно изменилась бы при правильном сэмплировании. Покажите, что в таком случае исходную задачу можно свести к правильной, если для объектов класса $+1$ ввести некоторый вес α. Найдите значение этого веса.

Задача 4 (2.5 балла). Пусть выборка $(x_i, y_i)_{i=1}^{\ell}$ генерируется из распределения $p(x, y)$ такого, что $\forall i = 1, ..., \ell$: $x_i \sim \text{Exp}(1)$, $y_i = f(x_i)$, где $f(z) = \sum\limits_{j=1}^{N} |z - j|$, где $N \in \mathbb{N}$ — фиксировано. Найдите смещение и разброс алгоритма $\forall x$: $\mu(X)(x) = \overline{X} = \dfrac{1}{\ell}\sum\limits_{i=1}^{\ell} x_i$ — среднее значение по обучающей выборке.