Введение в анализ данных
Контрольная работа
Пробный вариант

Задача 1 (3 балла). Ответьте на вопросы по линейным классификаторам и их обучению:

1. При обучении логистической регрессии решается задача

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\log\left(1+\exp(-y_i\langle w,x_i\rangle)\right)\to\min_w$$

   В формуле участвует $y_i\langle w,x_i\rangle$ — отступ на i-м объекте. Какой у него смысл, что означают его знак и абсолютное значение?

2. Каким должен быть отступ на объекте x_i, чтобы значение функции потерь $\log(1+\exp(-y_i\langle w,x_i\rangle))$ было минимальным?

3. Результатом обучения логистической регрессии является вектор весов w. Если нам дают объект x, как посчитать вероятность того, что он относится к положительному классу? Объясните все компоненты в формуле, которую запишете.

4. Логистическую регрессию можно обучать обычным или стохастическим градиентным спуском. Запишите формулу того, как выглядит шаг в обычном градиентном спуске и стохастическом градиентном спуске (в общем случае, не для логистической регрессии). Объясните все части этих формул. Опишите, в чём разница между ними, какие преимущества у стохастического градиентного спуска.

Задача 2 (3 балла). Ответьте на вопросы по обучению моделей:

1. Допустим, мы обучаем модель линейной регрессии на среднеквадратичную ошибку на некоторой выборке. Точно известно, что задача поиска оптимальных параметров модели имеет единственное решение. Мы рассматриваем два способа обучить эту модель: через аналитическую формулу для весов и через градиентный спуск. Чем отличаются эти два способа? Чем будут отличаться наборы весов, которые выдадут эти два метода?

2. Линейный классификатор имеет $(d+1)$ параметр, и мы знаем, что разделяющая поверхность в нём всегда линейна. Метод k ближайших соседей не имеет

обучаемых параметров, но разделяющая поверхность в нём может быть нелинейной и весьма сложной. Получаем некоторое противоречие: метод, который не подстраивает свои параметры под данные, оказывается сложнее. Как вы это объясните?

3. Говорят, что метрика accuracy чувствительна к балансу классов. Что это значит? Что можно сделать для корректной оценки работы модели в таком случае? Аргументируйте свой вариант для исправления.

Задача 3 (2 балла). Тамерлан решил запрограммировать градиентный спуск для своей задачи. Для начала он записал общую схему того, что будет программировать:

- Инициализация: $w^0 = 0$;
- Градиентный шаг:

$$w^t = w^{t-2} + \frac{1}{t^{10}} \nabla Q(w^{t-1});$$

- Останавливаемся, если $\|w^t\| < 1$.

После этого Тамерлан уехал в отпуск и поручил запрограммировать этот алгоритм своему двойнику Антону. Антону крайне не хочется самому разбираться в идеях Тамерлана. Помогите ему найти все эти ошибки, а заодно объясните, почему из-за них градиентный спуск будет работать не так, как надо. Подсказка: ошибок как минимум три.

Задача 4 (2 балла). Вам выдали классификатор $b(x)$, и вам предстоит разобраться, насколько она хороша. Для этого у вас есть тестовая выборка из 8 объектов. Ниже указаны правильные ответы и вероятности положительного класса от модели:

$b(x)$	0	0.2	0.3	0.4	0.5	0.7	0.9	0.95
y	-1	-1	-1	+1	+1	-1	+1	+1

Выполните следующие шаги:

1. Нарисуйте ROC-кривую и посчитайте AUC-ROC.
2. Посчитайте точность и полноту этой модели при пороге $t = 0.55$.
3. Можно ли достичь полноты в хотя бы 70% при точности в хотя бы 60%? Если да, укажите, при каком пороге.
4. Можно ли достичь полноты в 100% при точности в хотя бы 90%? Если да, укажите, при каком пороге.