Конспект лекции

А.Н. Шевляков

Машинное обучение

Задача классификации

# Классификация, kNN, кросс-валидация

## Метод kNN для задачи регрессии

Как «подкрутить» алгоритм kNN, чтобы он решал задачу регрессии?

Многие методы классификации (в т.ч. kNN) можно легко переделать для задачи регрессии (предсказания количественного признака).

Для классифицируемого объекта А находятся k его ближайших соседей по тренировочной выборке. Значение признака Y для А равно среднему значению признака Y его соседей. Оказывается, алгоритм классификации kNN можно легко переделать для нужд регресии.

## Методы выбора оптимальных параметров алгоритма. Кросс-валидация

Как найти оптимальное значение входного параметра алгоритма предсказания?

Часто алгоритм классификации (регрессии) зависит от значения входного параметра *p* (например, kNN зависит от значения параметра k).

Как найти оптимальное значение для *p*?

1) нужно взять как можно больше различных значений для *p*.

2) для каждого значения построить модель и проверить ее качество на тестовой выборке.

3) окончательно выбрать такое значение *p*, которое принадлежит модели с наилучшим качеством.

Недостатки: зависимость от конкретного разбиения на тренировочную и тестовую выборки.

**Кросс-валидация (cross-validation, скользящий контроль).** Модель обучается K раз на разных (K−1) подвыборках исходной выборки (белый цвет), а проверяется на одной подвыборке (каждый раз на разной, оранжевый цвет).  
Получаются K оценок качества модели, которые обычно усредняются, выдавая среднюю оценку качества классификации/регрессии на кросс-валидации.

Выбор параметров модели *p* с помощью кросс-валидации:

1) Нужно взять как можно больше различных значений для *p*.

2) Для каждого значения построить *K* моделей (для каждого разбиения данных при кросс-валидации)

3) Усреднить показатели качества этих *K* моделей (для каждой модели ее качество считается на её тестовой выборке).

4) Окончательно выбрать такое значение *p*, на котором достигается максимум усредненных показателей качества *K* моделей.

## План решения задачи классификации

Множество объектов разбить на 2 множества: тренировочную выборку Train и тестовую (проверочную) выборку Test.

Модель предсказания строится по объектам Train, а качество модели проверяется по объектам Test.

Классификацию можно свести к регрессии. Если предположить, что целевой признак Y числовой, то его можно найти с помощью модели регрессии.

Модель регрессии будет предсказывать вещественные числа (написаны в скобках), которые нужно будет округлять до 0 или 1.

Однако, как правило, с регрессией лучше не связываться, а сразу применять алгоритмы классификации.

**Бинарная классификация.** Если целевой признак бинарный (Y∈{0,1}), то классификация называется бинарной.

Далее вся теория будет касаться бинарной классификации (несложные обобщения для многоклассовой классификации предоставляются слушателям).

Качество классификации вычисляется по тестовой выборке. Для этого строят матрицу ошибок (confusion matrix).

Общая точность (accuracy): (TN+TP)/(TN+TP+FN+FP)

Но высокое значение accuracy еще не говорит о высоком качестве классификации!

Необходимо, чтобы precision и recall были как можно ближе к 1.