Классификация, kNN, кросс-валидация

Лекция 5

Постановка задачи предсказания (напоминалка)

Формальная постановка

Предсказание (prediction). Есть множество объектов M с известными значениями признака Y. Найти значение признака Y для нового объекта $A \notin M$. Y называется целевым признаком.

Предсказание значения номинального (категориального) признака Y называется задачей классификации.

	Рост	Вес	Пол
Вася	170	80	1
Даша	165	60	0
Маша	160	50	0
Петя	200	70	1

	Рост	Вес	Пол
A	180	75	?

План решения задачи классификации (для регрессии было аналогично)

Общий план решения задачи классификации

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

	Рост	Вес	Пол
Вася	170	80	1
Даша	165	60	0
Маша	160	50	0
Петя	200	70	1

Train	Рост	Вес	Пол
Вася	170	80	1
Даша	165	60	0

Test	Рост	Вес	Пол
Маша	160	50	0
Петя	200	70	1

Общий план решения задачи классификации

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

	Рост	Вес	Пол
Вася	170	80	1
Даша	165	60	0
Маша	160	50	0
Петя	200	70	1

Train	Рост	Вес	Пол
Вася	170	80	1
Даша	165	60	0

Test	Рост	Вес	Пол
Маша	160	50	0
Петя	200	70	1

Замечание о точности классификации (дежавю)

Модель классификации не обязана давать точный ответ на объектах тренировочной (!) выборки (которые использовались при построении модели).

Замечание о точности регрессии

То есть если модель предсказания была построена по данным

Train	Рост	Вес	Пол(Y)
Вася	170	80	1
Даша	165	60	0

а вы решили приколоться и подать ей на вход объект

Joke	Рост	Вес	Пол(Y)
Даша	165	60	?

то модель не обязана выдавать вам 0 в качестве ответа.

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.

Test	Рост	Вес	Пол
Маша	160	50	0(0.55)
Петя	200	70	1(0.81)

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

Оценивание качества модели по тестовой выборке

Допустим по тренировочной выборке мы научились предсказывать целевой признак Y. Как оценить качество предсказаний по тестовой выборке?
(В таблице в скобках указаны предсказанные значения.)

Test	Рост	Вес	Пол
Маша	160	50	0(0)
Петя	200	70	1(0)

Бинарная классификация

Если целевой признак бинарный ($Y \in \{0,1\}$), то классификация называется бинарной.

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

Критерии качества классификации

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

		Истинный класс	
		0	1
Предсказанный класс	0	TN	FN
	1	FP	TP

TN=true positive FN=false negative
TP=true negative FP=false positive

Пример: медицинские анализы

Анализы – это простейшие классификаторы, которые «предсказывают» 0 (человек здоров) или 1 (болен).

На следующих слайдах будет показана матрица ошибок анализа на ВИЧ, которому была подана тестовая выборка из 10000 американских белых мужчин не употребляющих наркотики (1989г).

Свойства анализа

Отсюда, кстати, можно посчитать вероятность наличия болезни у пациента при условии, что анализ казался положительным (чему она равна?).

		Истинный класс	
		0	1
Предсказанный класс	0	9989	0
	1	10	1

Невежество врачей и пациентов (а также юристов, журналистов и др.), связанное с теорией вероятностей, обсуждается в книге Л.Млодинов «(Не)совершенная случайность».

Критерии качества классификации

Интуитивно понятно, чем больше числа на диагонали, тем лучше, а FN и FP должны быть =0. Но на практике уменьшение FN приводит к увеличению FP (и наоборот).

		Истинный класс	
		0	1
Предсказанный класс	0	TN	FN
	1	FP	TP

Критерии качества классификации

1.Общая точность (accuracy):

$$(TN+TP)/(TN+TP+FN+FP)$$

		Истинный класс	
		0	1
Предсказанный класс	0	TN	FN
	1	FP	TP

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

Проблема accuracy

Классификатор

		Истинный класс	
		0	1
Предсказанный класс	0	9990	10
	1	0	0

имеет значение accuracy=0.999 но фактически такая классификация бесполезна (особенно когда принадлежность классу 1 гораздо важнее чем классу 0). Accuracy бесполезна, если один из классов гораздо больше другого.

Вопросик про accuracy

Правда, что очень легко построить классификатор с accuracy не меньше 0.5?

		Истинный класс	
		0	1
Предсказанный класс	0	TN	FN
	1	FP	TP

Потому что если accuracy вашего классификатора <0.5 то можно…

Вопросик про accuracy

Правда, что очень легко построить классификатор с accuracy не меньше 0.5?

		Истинный класс	
		0	1
Предсказанный класс	0	TN	FN
	1	FP	TP

Потому что если accuracy вашего классификатора <0.5 то можно инвертировать его ответы и получить требуемую accuracy.

Критерии качества классификации

2. точность (precision):

$$TP/(TP+FP)$$

3. полнота (recall):

$$TP/(TP+FN)$$

		Истинный класс	
		0	1
Предсказанный класс	0	TN	FN
	1	FP	TP

Желательно, чтобы обе эти характеристики были близки к 1.

Проблема accuracy

Для анализа на ВИЧ имеем следующие величины:

		Истинный класс	
		0	1
Предсказанный класс	0	9989	0
	1	10	1

precision=1/(1+10)=0.09
recall=1/(1+0)=1

Критерии качества классификации

4. F-value (корректно агрегирует точность и полноту в одно выражение):

$$F=2*precision*recall/(precision+recall)$$

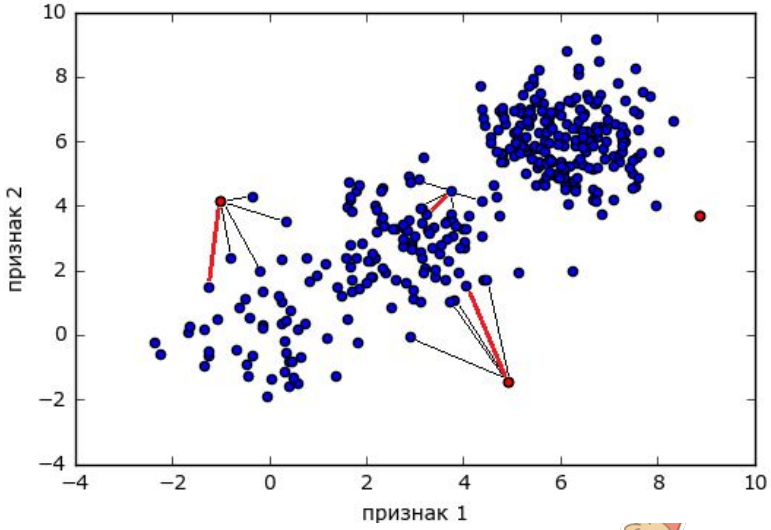
Это позволяет максимизировать одну величину (F-value), вместо максимизации precision и recall. F-value для анализа на ВИЧ равна

$$F=2*0.09*1/(0.09+1)=0.18/1.09=0.16$$

Метод k ближайших соседей (kNN)

Основная суть метода kNN

kNN – метрический метод классификации, то есть объекты представляются в виде точек в пространстве и между ними вычисляются расстояния.
Следовательно, признаки должны быть нормированы (приведены к одному масштабу)
Число k – входной параметр алгоритма и может быть оптимально настроен.

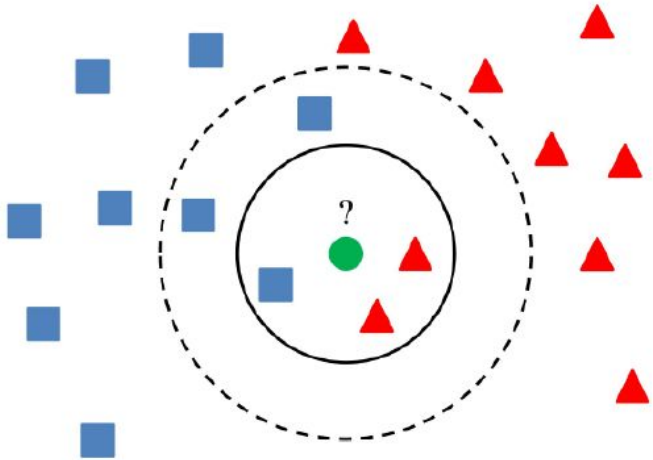


Правило классификации

Для классифицируемого объекта *A* находятся *k* его ближайших соседей по тренировочной выборке.
Объект *A* относится к тому классу, который является наиболее распространённым среди *k* соседей.

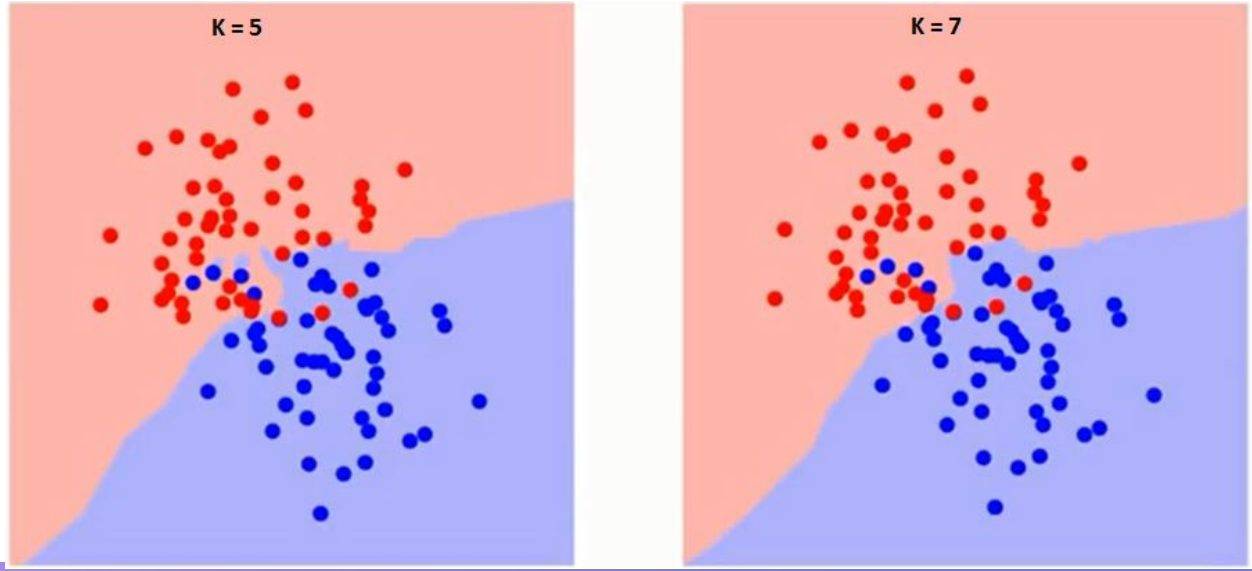
Зависимость результата от параметра k

На картинке при k=3 зелёный объект будет отнесен к «красному» классу, а при k=5 зелёный объект будет отнесен к «синему» классу.
Для четного k результат классификации может быть не определен.



Фактически всё пространство разбивается на зоны

Классифицируемый объект, попавший в зону, будет отнесен к соответствующему классу.



Экстремальные значения параметра k

k=1 (ближайший сосед)
k=|Train| (объем тренировочной выборки). Как вы думаете, какое значение будет выдавать алгоритм?

Очевидно, оптимальное значение для k где-то между 1 и |Train|. На этом значении достигается минимум ошибки на тестовой выборке (способы выбора оптимального значения параметра будет в конце лекции).

Вы можете предложить модификации метода kNN

1. Выбор метрики (можете использовать самые экзотические).
2. …

Вы можете предложить модификации метода kNN

1. Выбор метрики (можете использовать самые экзотические).
2. Соседей можно взвесить. Класс наиболее близких соседей приобретает больший вес при принятии решения (это, кстати, решает проблему четных k).

Проблемы метода kNN

1. Неустойчивость к выбросам.
2. Как правило, плохо работает, когда признаков очень много.
3. С помощью одного лишь kNN сложные задачи (как правило) не решить, но его kNN часто используется для построения мета-признаков (прогноз kNN подается на вход более сложным моделям).

Метод kNN для задачи регрессии

kNN для задачи регрессии

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

Для классифицируемого объекта A находятся k его ближайших соседей по тренировочной выборке. Значение признака Y для A равно…

kNN для задачи регрессии

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

Для классифицируемого объекта A находятся k его ближайших соседей по тренировочной выборке. Значение признака Y для A равно среднему значению признака Y его соседей.

Методы выбора оптимальных параметров алгоритма. Кросс-валидация

Часто алгоритм классификации (регрессии)

Часто алгоритм классификации (регрессии) зависит от значения входного параметра *p* (например, kNN зависит от значения параметра k).
Как найти оптимальное значение для *p*?

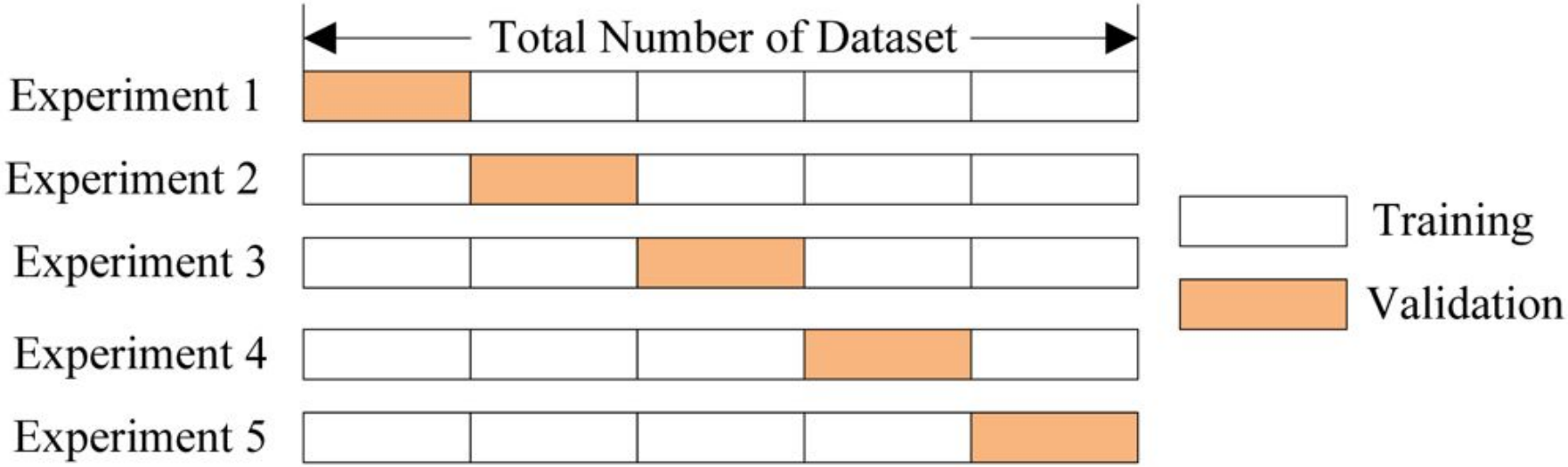
Самый простой способ выбора параметра

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)

Нужно разделить всю выборку на K частей (на рисунке K=5 – и такое значение часто берут).

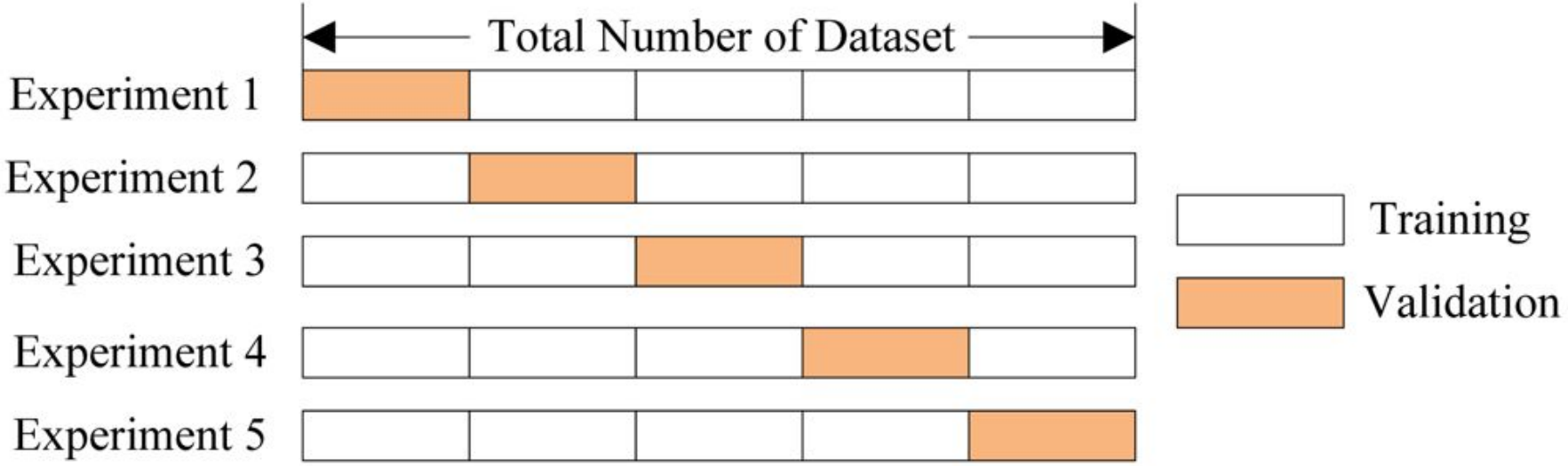


Кросс-валидация (cross-validation, скользящий контроль)

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.

Экстремальный случай кросс-валидации

Если K=«количество всех объектов», то получаем схему leave-one-out (тестовая выборка при каждом запуске будет состоять ровно из 1 объекта!!!).



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.

Использованная литература

1. https://habrahabr.ru/company/ods/blog/328372/ (про критерии качества классификации)
2. https://habrahabr.ru/company/ods/blog/322534/ (про kNN)
3. Т.Сегаран «Программируем коллективный разум» (определение стоимости вина с помощью kNN)