

Классификация данных. Основные понятия

Курс «Интеллектуальные информационные системы»
Кафедра управления и интеллектуальных технологий
НИУ «МЭИ»
Осень 2021 г.

Задача классификации

Задача классификации – отнести новый объект к одному из заранее определенных классов на основе некоторой функции (алгоритма, решающего правила, классификатора)



Виды классификации:

- Бинарная классификация (классификация на 2 класса, $k=2$)
- На k непересекающихся классов ($k>2$)
- На k классов, которые могут пересекаться

Меры близости и расстояния

Евклидово расстояние

$$d(\vec{X}_j, \vec{X}_l) = \sqrt{\sum_{i=1}^M (x_j^{(i)} - x_l^{(i)})^2}$$

Расстояние городских кварталов

$$d(\vec{X}_j, \vec{X}_l) = \sum_{i=1}^M |x_j^{(i)} - x_l^{(i)}|$$

Косинусоидальная мера близости.
Показывает косинус угла между векторами.

Стремится к 1, когда объекты похожи между собой

$$d(\vec{X}_j, \vec{X}_l) = \cos(\vec{X}_j, \vec{X}_l) = \frac{\sum_{i=1}^M x_j^{(i)} x_l^{(i)}}{\sqrt{\sum_{i=1}^M (x_j^{(i)})^2 \sum_{i=1}^M (x_l^{(i)})^2}}$$

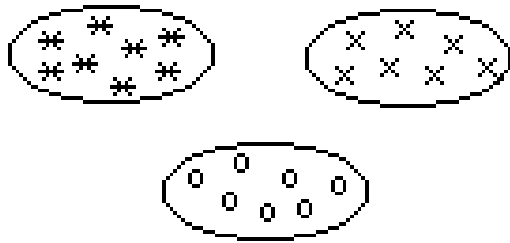
Формирование выборок

- Эффективность методов Machine Learning сильно зависит от того, как были сформированы выборки.
- Выборки должны быть:
 - Независимо извлеченными из генеральной совокупности
 - Представительными (репрезентативными)
 - Содержать минимум нетипичных объектов
- Не так важно, как выглядит генеральная совокупность во всем пространстве признаков. Гораздо важнее, как она выглядит в районе границы между двумя классами

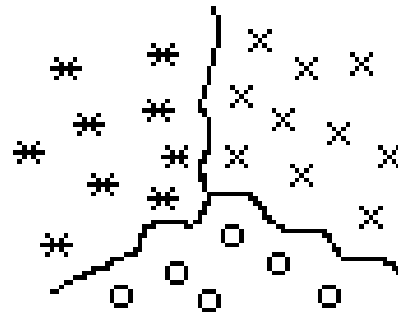
Неидеальность разметки объектов – разные эксперты могут отнести один объект к разным классам. Как поступать?

Как оценить выборку?

Ядерная (центроидная) модель



Модель рассеяния



Модель засорения



Средняя сумма внутриклассовой дисперсии:

$$Q_1 = \frac{1}{N_k} \sum_{j=1}^{N_k} d^2(\vec{X}_j, \vec{X}_k) \quad \rightarrow \quad Q_1^* = \frac{1}{M} \sum_{k=1}^M \frac{1}{N_k} \sum_{j=1}^{N_k} d^2(\vec{X}_j, \vec{X}_k),$$

Средняя сумма квадратов внутриклассовых попарных расстояний

$$Q_2 = \frac{1}{N_k} \sum_{l=1}^{N_k} \sum_{j=1, j \neq l}^{N_k} d^2(\vec{X}_l, \vec{X}_j) \quad \rightarrow \quad Q_2^* = \frac{1}{M} \sum_{k=1}^M \frac{1}{N_k} \sum_{l=1}^{N_k} \sum_{j=1, j \neq l}^{N_k} d^2(\vec{X}_l, \vec{X}_j)$$

Как оценить выборку? (2)

Средняя сумма квадратов
межклассовых попарных
расстояний

$$Q_3 = \frac{1}{N_k N_s} \sum_{l=1}^{N_k} \sum_{j=1, j \neq l}^{N_s} d^2(\vec{X}_l, \vec{X}_j) \quad \text{или} \quad Q_3 = \frac{1}{M} \sum_{k=1}^M \sum_{s=1, s \neq k}^M \frac{1}{N_k N_s} \sum_{l=1}^{N_k} \sum_{j=1, j \neq l}^{N_s} d^2(\vec{X}_l, \vec{X}_j)$$

Обобщенный функционал

$$Q_4 = \frac{Q_3}{Q_2}$$

На основе такого анализа исследователь может: 1) объединить несколько близких небольших классов в один; 2) удалить “нехарактерные” шумовые элементы, расположенные вдалеке от центра классов (модель засорения); 3) заново сформировать выборку, увеличив (уменьшив) количество классов или количество элементов.

Свойства сформированных выборок

- любая обучающая выборка конечного размера не является полной, т.е. не содержит необходимого количества элементов для проведения безошибочной классификации;
- элементы обучающей выборки обычно имеют произвольное распределение в пространстве признаков и, как следствие, получаемые решающие правила могут обладать неодинаковой достоверностью в различных областях изменения параметров;
- выборки, как правило, содержат шумовые (нерелевантные, не относящиеся к указанным классам) элементы, другую противоречивую или ошибочную информацию, которая так или иначе попадает в обучающую выборку.

Виды выборки

Часть размеченных документов оставляют для обучения, часть – для оценки точности метода.

Выборка (dataset) делится на две части:

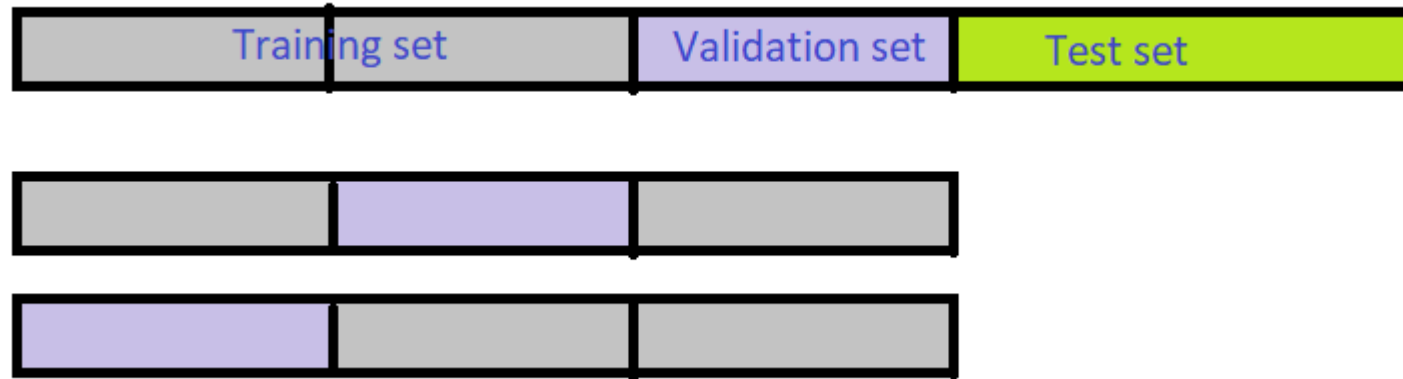
- Обучающая (тренировочная, training set)
- Тестовая (экзаменационная, test set)

Нобуч > Nтест(обычно 70/30)



Оценка качества классификации в задачах Data Mining

- Оценка качества с помощью k -кратной перекрестной проверки (k -fold cross validation)



- Оценка качества с помощью скользящего контроля (или «метод складного ножа», «Jackknife») – для небольших выборок
- Bootstrap – имитация статистического выбора. Суть метода заключается в формировании множества выборок на основе случайного выбора с повторениями.

Оценка качества классификации в задачах Text Mining (2)

Ошибка классификации – несовпадение метки, назначенной классификатором с меткой, назначенной экспертом (учителем).

Точность (правильность, аккуратность)

$$\text{Accuracy} = \frac{P}{N}$$

P- количество объектов, по которым классификатор принял правильное решение

$$\text{Точность Precision} = \frac{TP}{TP+FP}$$

$$\text{Полнота Recall} = \frac{TP}{TP+FN}$$

$$\text{F-measure} = \frac{2(\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}}$$

Матрица неточностей (Confusion matrix)

	Оценка эксперта	
	Положительная	Отрицательная
Оценка классификатора		
Положительная	TP	FP
Отрицательная	FN	TN

Матрица неточностей

$$\text{Precision}_k = \frac{A_{kk}}{\sum_{i=1}^n A_{ki}}$$

$$\text{Recall}_k = \frac{A_{kk}}{\sum_{i=1}^n A_{ik}}$$

Точность классификатора =
среднее его точности по всем
классам k (macro average)

Micro average Precision:

$$\frac{\sum_{i=1}^n TP_i}{\sum_{i=1}^n TP_i + \sum_{i=1}^n FN_i}$$

Аналогично – другие показатели по micro-ес

	0.91	0.96	0.94	0.75	1.00	0.83	0.85	0.97	1.00	0.86	1.00	0.79	1.00	0.75	1.00	1.00	0.96	0.90	0.81	0.89	0.94	0.98	0.86	0.89	0.94	0.92	0.96
0.80		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
0.95	1	94	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
1.00	2	0	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.29	3	0	0	6	0	0	3	2	0	1	0	0	0	0	0	0	1	1	0	0	1	0	1	3	0	2	0
1.00	4	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.50	5	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	1	1
0.92	6	1	0	0	0	0	152	0	0	1	0	0	0	0	0	0	0	1	4	2	3	0	0	0	0	2	0
0.97	7	1	0	1	0	0	0	256	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	2	0
0.33	8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
0.97	9	0	0	0	0	0	0	0	0	69	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
0.82	10	0	0	0	0	0	2	0	0	0	18	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
0.87	11	0	0	0	0	0	0	0	0	0	0	34	0	4	0	0	0	0	0	0	0	0	0	1	0	0	0
1.00	12	0	0	0	0	0	0	0	0	0	0	0	37	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.57	13	0	0	0	0	0	0	0	0	0	0	9	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0
0.63	14	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	3	0	0	0	0	0	0	0	0	0
0.50	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	1	1	0	0	0	0	0	0
0.77	16	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	47	0	1	3	4	0	0	2	0	1	0
0.87	17	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	69	1	2	5	0	0	0	0	0	0
0.97	18	0	0	0	0	1	4	0	0	1	0	0	0	0	0	0	0	0	197	1	0	0	0	0	0	0	0
0.78	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	35	183	13	0	0	2	0	1	0	
0.97	20	0	0	0	0	0	10	3	0	1	0	0	0	0	0	0	0	0	4	702	0	0	0	0	6	0	
0.93	21	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	56	0	2	0	0	0
0.29	22	0	0	1	0	0	2	0	0	6	0	0	0	0	0	0	0	1	1	1	0	6	2	0	1	0	
0.91	23	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	3	6	0	0	115	0	0	0
1.00	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	
0.93	25	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2	4	5	0	0	0	1	196	0
0.98	26	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	78

Несбалансированная выборка

Несбалансированные выборки

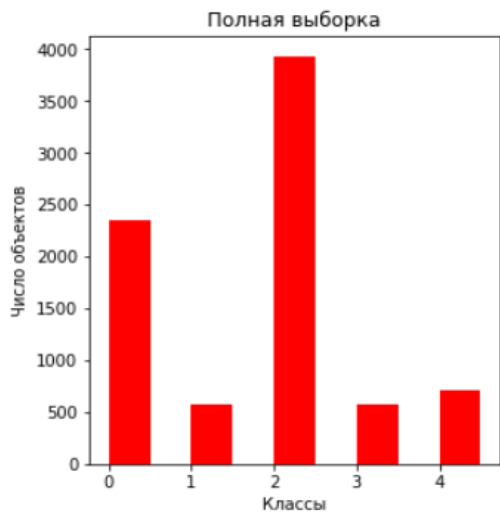
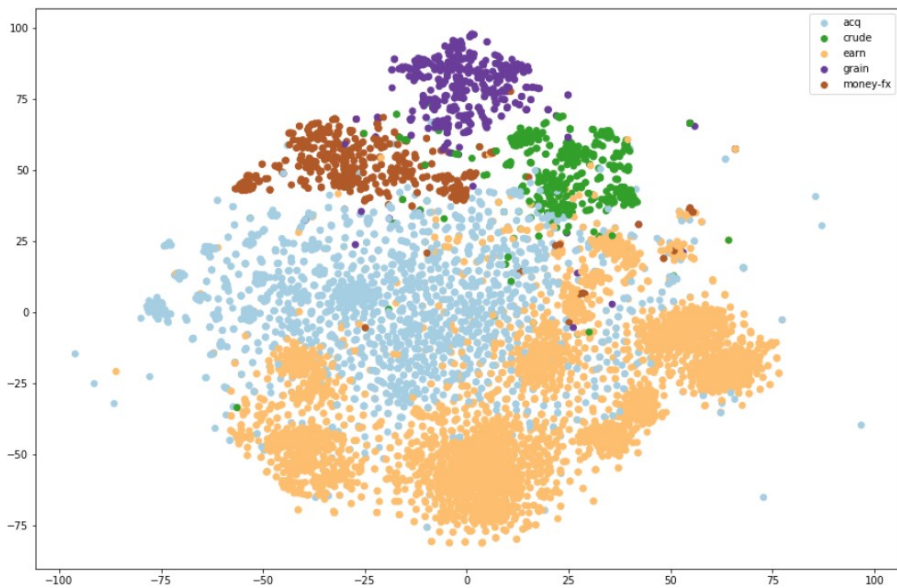
Несбалансированная выборка (Unbalanced sample) – количество объектов каждого класса очень сильно разнится.

- Низкая точность классификации у малых классов
- Часто классификатору бывает выгодно объекты всех малых классов приписать к самому большому и не обучаться на малых.

Нужно приводить к сбалансированному виду –

- Oversampling – дублирование объектов малых классов
- Undersampling – удаление объектов из больших классов
- SMOTE (Synthetic Minority Over-sampling Technique) – создание искусственных объектов малых классов

Несбалансированные выборки



Несбалансированные выборки (Imbalanced Data) – выборки, в которых количество объектов в классах очень сильно разнится.

Классы, большие по объему – мажоритарные, классы малых объемов – миноритарные.

Следствия:

- Неравномерное обучение на разных классах
- Плохое распознавание объектов малых классов
- Нельзя использовать Accuracy
- Выгоднее отнести все объекты к мажоритарному классу

- K1 = 10 объектов
- K2 = 100 объектов

$$\text{Accuracy} = \frac{TP+TN}{N}$$

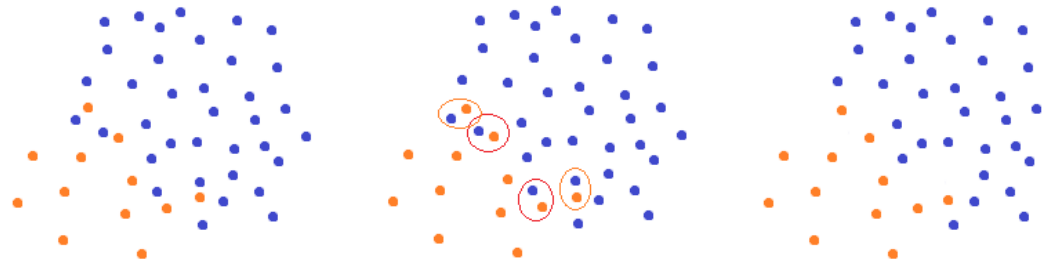
Эксперт:	K1	K2
Модель:		
K1	6	16
K2	4	84

Oversampling

- Random Over Samper
- SMOTE – Находит n ближайших соседей, для каждого миноритарного объекта. «Соединяет» их линиями и случайным образом создает новые объекты на этих линиях.
- ADASYN - вместо линейной генерации объектов, добавляется некоторый шум, позволяющий создать нелинейную зависимость между родителем и синтезированным объектом

Undersampling

- Random Under Sampler
- Tomek Links - Пара (E_i, E_j) называется связью Томека, если
 - $$\begin{cases} d(E_i, E_l) < d(E_i, E_j) \\ d(E_j, E_l) < d(E_i, E_j) \end{cases}$$



Объекты мажоритарного класса, входящие в связи Томека удаляются

- EditedNearestNeighbors - Объект мажоритарного класса удаляется при условии, если не все его k ближайших соседей имеют метку мажоритарного класса.

AUC ROC

ROC - receiver operating characteristic, кривая ошибок

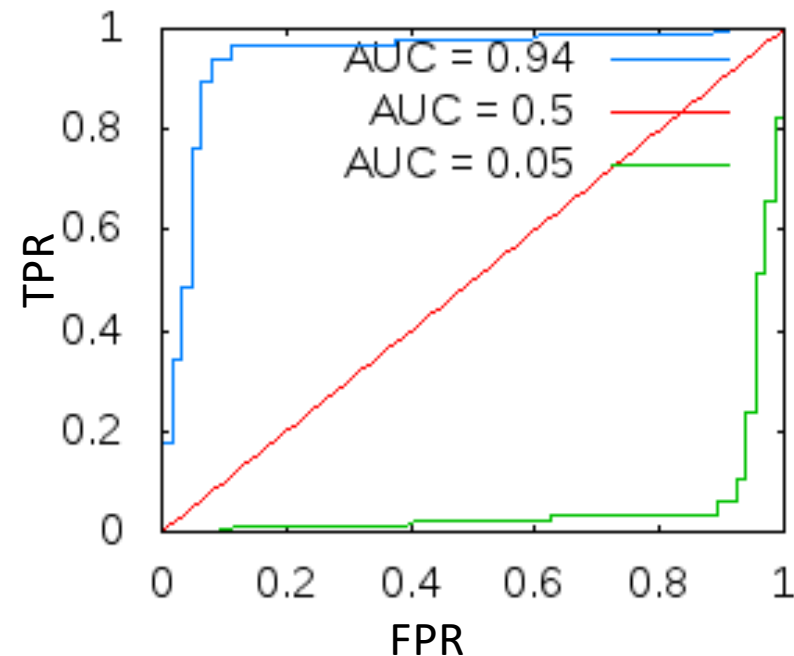
AUC ROC - площадь под кривой ошибок, Area Under ROC Curve –
Зависимость доли верных положительных классификаций от
доли ложных положительных классификаций при варьировании
порога решающего правила.

AUC ROC — эквивалентна вероятности, что классификатор
присвоит большее значение случайно выбранному
позитивному объекту, чем случайно выбранному негативному
объекту.

Когда **AUC = 0.5**, то данный классификатор равен случайному.

Если **AUC < 0.5**, то можно просто перевернуть выдаваемые
значения классификатором.

Визуально - чем больше график прижимается к верхнему
левому углу, тем больше значение AUC



$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{TN + FP}$$

Переобучение

Качество на обучающей выборке

Ассурасу $a = 0.85$

Ассурасу $a' = 0.99$

} Какой алгоритм лучше?

Проверяем качество на тестовой выборке, которая не участвовала в процессе обучения и настройки параметров:

Ассурасу $a = 0.84$

Ассурасу $a' = 0.79$

Переобучение (overfitting, overtraining) – нежелательное явление, когда качество классификации на тестовой выборке (или на реальных данных) существенно ниже, чем на обучающей выборке.

