Классификация данных. Основные понятия

Курс «Интеллектуальные информационные системы»

Кафедра управления и информатики НИУ «МЭИ»

Осень 2018 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 документы похожи между собой

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

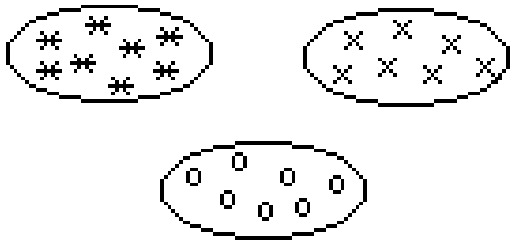
Формирование обучающих и экзаменационных выборок

- Эффективность методов Machine Learning сильно зависит от того, как были сформированы обучающие 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

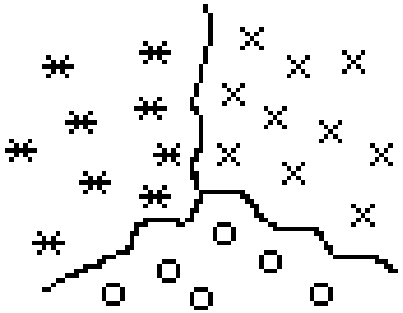
Неидеальность разметки документов – разные эксперты могут отнести документ к разным классам. Как поступать?

Как оценить выборку?

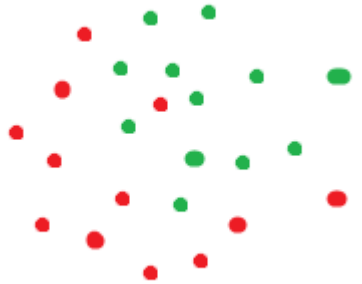
Ядерная (центроидная) модель



Модель рассеяния



Модель засорения



Средняя сумма внутриклассовой дисперсии:

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

Средняя сумма квадратов внутриклассовых попарных расстояний

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

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.

Оценка точности классификации в задачах Data Mining

Часть размеченных документов оставляют для обучения, часть – для оценки точности метода. Обычно используют следующие методы оценки:

- *Оценка точности по экзаменационным выборкам.* $N_{обуч} > N_{экзам}$
- *Оценка точности с помощью скользящего контроля (или «метод складного ножа», «Jackknife») – для небольших выборок*
- *Оценка точности с помощью k–кратной перекрестной проверки (k–fold cross validation)*
- Bootstrap – имитация статистического выбора. Суть метода заключается в формировании множества выборок на основе случайного выбора с повторениями.

Оценка точность классификации в задачах Text Mining (2)

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

Точность (правильность, аккуратность)

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

P- количество документов, по которым классификатор принял правильное решение

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

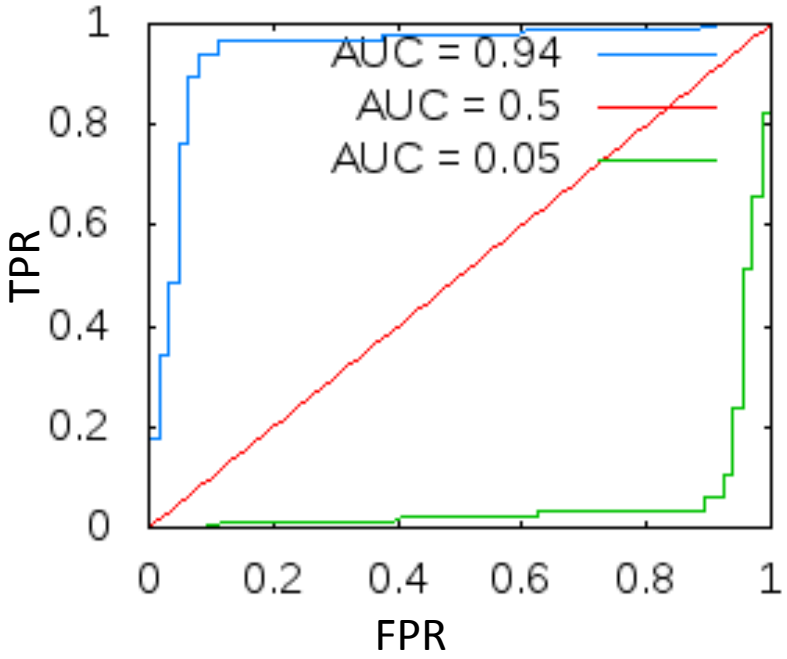
Полнота Recall $= \frac{TP}{TP+FN}$

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

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

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

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

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

AUC ROC (2)

id	оценка	класс
1	0.5	0
2	0.1	0
3	0.2	0
4	0.6	1
5	0.2	1
6	0.3	1
7	0.0	0

Табл. 1

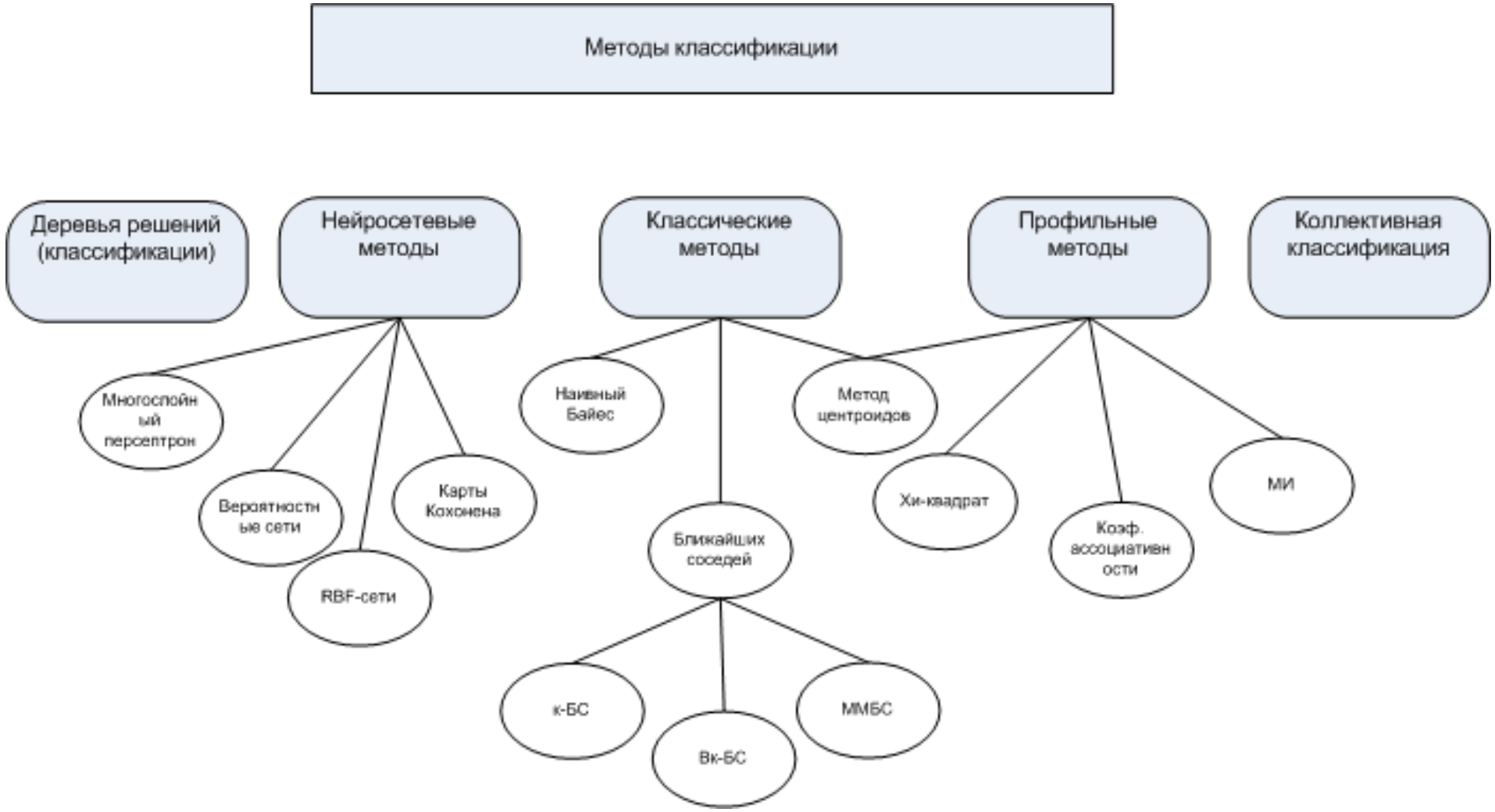
id	оценка	класс
4	0.6	1
1	0.5	0
6	0.3	1
3	0.2	0
5	0.2	1
2	0.1	0
7	0.0	0

Табл. 2

id	> 0.25	класс
4	1	1
1	1	0
6	1	1
3	0	0
5	0	1
2	0	0
7	0	0

Табл. 3

Систематизация методов классификации

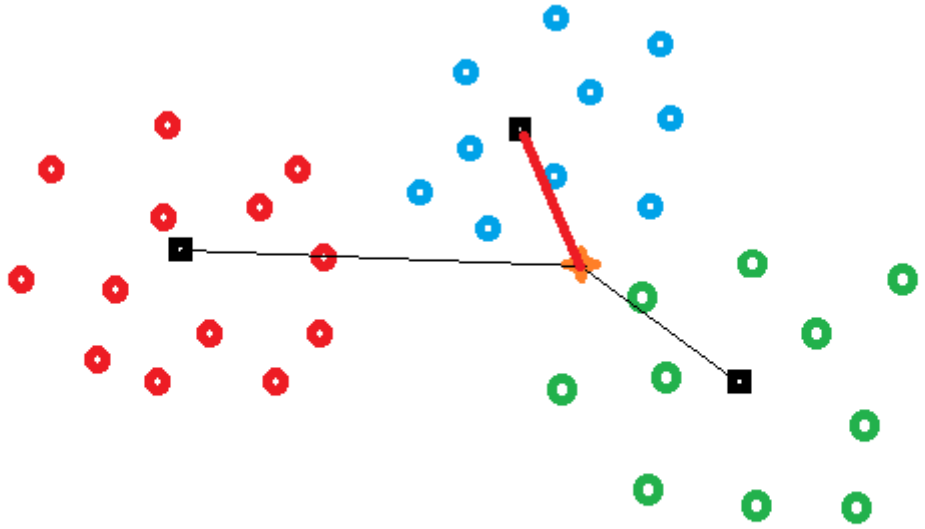


Центроидный метод

Центроид – вектор со средними значениями весов терминов документов данного класса. «Центр тяжести».
Классифицируемый объект относится к классу с наиболее близким центроидом.

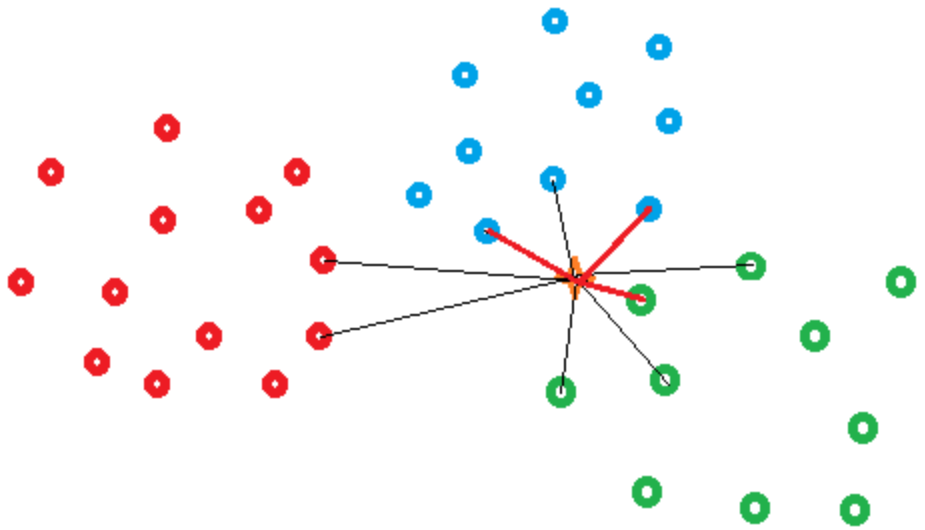
$$\vec{\mathbf{C}}_k = \frac{1}{N_k}\sum_{j=1}^{N_k}\vec{X}_j$$

Роккио: $$\vec{C}_k = \alpha\frac{1}{N_k}\sum_{j=1}^{N_k}\cos(\vec{C}_k,\vec{X}_j) - \frac{\beta}{N-N_k}\sum_{l=1}^{N-N_k}\cos(\vec{C}_k,\vec{X}_l)$$



Правило ближайшего соседа (БС)

Классифицируемый объект относится к тому классу, к которому относится ближайший к нему сосед.



Семейство методов БС

- кБС – Решение принимается на основании анализа к ближайших соседей. Обычно к - нечетное число [5;25]
- Взвешенный кБС – наиболее близкие соседи имеют больший вес при голосовании.
- Модифицированный МБС – поиск соседей только определенной области признакового пространства, с целью сокращения вычислительных операций.