

Лекция 3

ЗАДАЧА КЛАССИФИКАЦИИ

Задача классификации

Классификация - системное распределение изучаемых предметов, явлений, процессов по родам, видам, типам, по каким-либо существенным признакам для удобства их исследования; группировка исходных понятий и расположение их в определенном порядке, отражающем степень этого сходства.

Классификация - упорядоченное по некоторому принципу множество объектов, которые имеют сходные классификационные признаки (одно или несколько свойств), выбранных для определения сходства или различия между этими объектами.

Классификация требует соблюдения следующих правил:

- в каждом акте деления необходимо применять только одно основание;
- деление должно быть соразмерным, т.е. общий объем видовых понятий должен равняться объему делимого родового понятия;
- члены деления должны взаимно исключать друг друга, их объемы не должны перекрещиваться;
- деление должно быть последовательным.

Задача классификации

- *Классификация* - это закономерность, позволяющая делать вывод относительно определения характеристик конкретной группы.
- Для проведения *классификации* должны присутствовать признаки, характеризующие группу, к которой принадлежит то или иное событие или объект (обычно при этом на основании анализа уже классифицированных событий формулируются некие правила).
- *Классификация* относится к стратегии *обучения с учителем* (*supervised learning*), которое также именуют контролируемым или управляемым обучением.
- Задачей *классификации* часто называют предсказание категориальной зависимой переменной (т.е. зависимой переменной, являющейся категорией) на основе выборки непрерывных и/или категориальных переменных.

Постановка задачи

Предполагается, что уже имеется какое-то количество n объектов, для каждого из которых известен некоторый набор из m признаков (факторов) и номер класса, к которому этот объект принадлежит, т.е. сырье данные, используемые для решения задачи классификации, имеют вид:

Номер наблюдения, i	Значения факторов			Значения переменной отклика (номер класса)
1	$x_{1,1}$...	$x_{1,m}$	y_1
...
i	$x_{i,1}$...	$x_{i,m}$	y_i
...
n	$x_{n,1}$...	$x_{n,m}$	y_n

Здесь значения переменной отклика – номер класса, которому принадлежит объект, т.е. $y_i \in \{1, \dots, K\}$, для всех $i = 1, \dots, n$, K – (известное) количество классов

Постановка задачи

Как и в задаче регрессионного анализа, предположим, что имеется n объектов, каждый из которых описывается m признаками.

Будем нумеровать объекты индексом i ($i = 1, \dots, n$), а признаки (значения которых могут быть получены непосредственным измерением) – индексом j ($j = 1, \dots, m$).

Для объекта с номером i обозначим через $x_{i,j}$ – значения признака j ; y_i – значение зависимого признака объекта i .

Пример. Пусть объекты – это клиенты банка, наблюдаемый признак x – уровень их заработной платы, прогнозируемый признак y – состояние кредитной карты. Цель исследования – спрогнозировать, «уйдёт ли в минус» тот или иной клиент банка (владелец банковской карты).

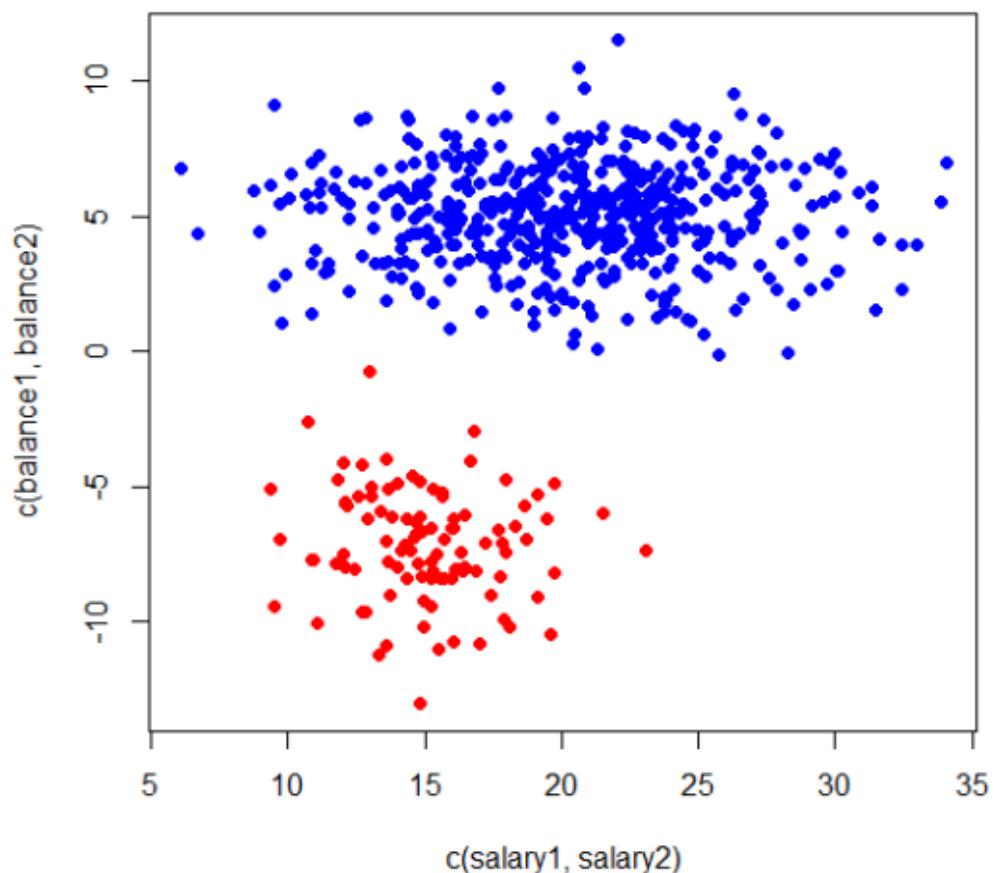
В этом примере $m = 1$. Данные n обследованных клиентов запишем как пары (x_i, y_i) , $i = 1, \dots, n$. Каждой такой паре можем поставить в соответствие точку на координатной плоскости.

«Разобьём» всех клиентов на два класса:

- «благонадёжные» (т.е. имеющие неотрицательный баланс на карте)
- «неблагонадёжные» (т.е. имеющие отрицательный баланс на карте).

Постановка задачи. Пример

Зависимость между зарплатой и балансом карты



Синие точки соответствуют клиентам банка, имеющим неотрицательный баланс кредитной карты, красным – отрицательный.

Одни и те же значения по оси абсцисс могут соответствовать как синим, так и красным точкам.

Однако можно заметить, что большим значениям признака x (заработной плате) соответствует большее число синих точек, чем красных.

Виды классификации

Вспомогательная

(искусственная) классификация:
производится по внешнему
признаку и служит для придания
множеству предметов
(процессов, явлений) нужного
порядка;

Естественная классификация:

производится по существенным
признакам, характеризующим
внутреннюю общность предметов и
явлений, предполагает и закрепляет
результаты изучения закономерностей
классифицируемых объектов.

В зависимости от выбранных признаков, их сочетания и процедуры
деления понятий **классификация** может быть:

Простая - деление родового
понятия только по признаку и
только один раз до раскрытия
всех видов.

Сложная - применяется для
деления одного понятия по
разным основаниям и синтеза
таких простых делений в единое
целое.

Классификация может быть **одномерной** (по одному признаку)
и **многомерной** (по двум и более признакам).

Процесс классификации

1 этап. Конструирование модели: описание множества предопределенных классов.

- Каждый пример набора данных относится к одному предопределенному классу.
- На этом этапе используется обучающее множество, на нем происходит конструирование модели.
- Полученная модель представлена классификационными правилами, деревом решений или математической формулой.

2 этап. Использование модели: классификация новых или неизвестных значений.

- Оценка правильности (точности) модели.
 1. Известные значения из тестового примера сравниваются с результатами использования полученной модели.
 2. Уровень точности - процент правильно классифицированных примеров в тестовом множестве.
 3. Тестовое множество, т.е. множество, на котором тестируется построенная модель, не должно зависеть от обучающего множества.
- Если точность модели допустима, возможно использование модели для классификации новых примеров, класс которых неизвестен.

Основные методы классификации:

классификация с помощью деревьев решений;

байесовская (наивная) классификация;

классификация при помощи искусственных нейронных сетей;

классификация методом опорных векторов;

статистические методы, в частности, линейная регрессия;

классификация при помощи метода ближайшего соседа;

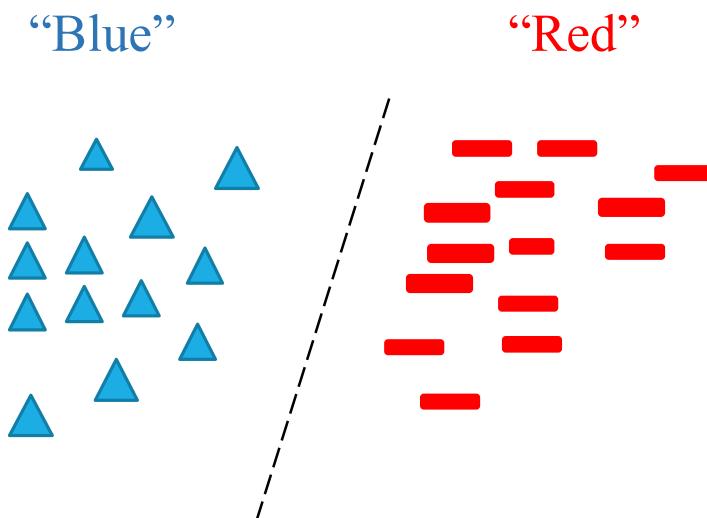
классификация CBR-методом;

классификация при помощи генетических алгоритмов.

Метод опорных векторов

Support Vector Machine - SVM

- относится к группе **границых методов** (классы определяются при помощи границ областей);
- Назначение** - с помощью SVM решают задачи **бинарной классификации**;
- в основе метода** - понятие **плоскостей решений** (плоскость решения разделяет объекты с разной классовой принадлежностью).



Разделение классов прямой линией

Разделяющая линия задает границу, справа от которой - все объекты типа “blue”, слева - типа “red”.

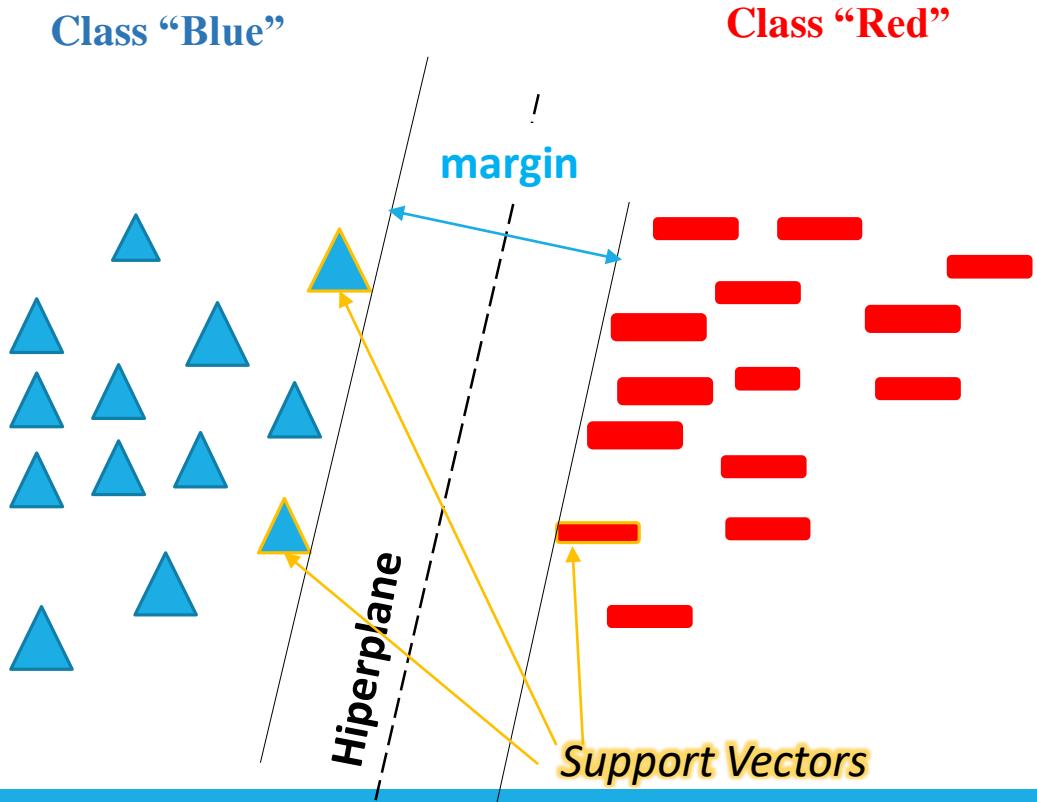
Новый объект, попадающий направо, классифицируется как объект класса **blue** или - как объект класса **red**, если он расположился по левую сторону от разделяющей прямой.

В этом случае
каждый объект характеризуется двумя
измерениями.

Метод опорных векторов

Цель метода опорных векторов - найти плоскость (*гиперплоскость*), разделяющую два множества объектов.

Метод отыскивает образцы, находящиеся на границах между двумя классами, т.е. **опорные векторы**.



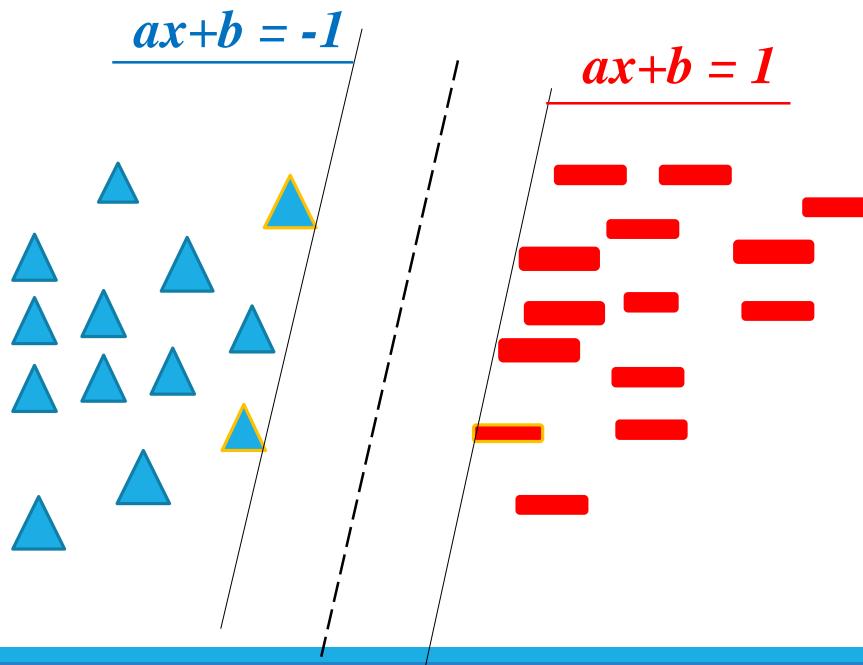
Опорными векторами называются объекты множества, лежащие на границах областей.

Классификация считается хорошей, если область между границами пуста.

Линейный метод опорных векторов (SVM)

Решение задачи **бинарной классификации** при помощи **метода опорных векторов** заключается в поиске некоторой линейной функции, которая правильно разделяет набор данных на два класса.

Рассмотрим задачу классификации, где число классов равно двум: поиск функции $f(x)$, принимающей значения **меньше нуля** для векторов **одного класса** и **больше нуля** - для векторов **другого класса**.



Дан тренировочный набор векторов пространства, для которых известна их принадлежность к одному из классов.

Семейство классифицирующих функций можно описать через функцию $f(x)$.

Гиперплоскость определена вектором a и значением b , т.е. $f(x)=ax+b$.

Линейный метод опорных векторов (SVM)

- В результате решения задачи (построения SVM-модели) будет найдена функция, принимающая значения меньше нуля для векторов одного класса и больше нуля - для векторов другого класса.
- Для каждого нового объекта отрицательное или положительное значение определяет принадлежность объекта к одному из классов.
- Наилучшей функцией классификации является функция, для которой ожидаемый риск минимален. Понятие **ожидаемого риска** в данном случае означает ожидаемый уровень ошибки классификации.
- Напрямую оценить ожидаемый уровень ошибки построенной модели невозможно, это можно сделать при помощи понятия эмпирического риска.
- Но следует учитывать, что минимизация эмпирического риска не всегда приводит к минимизации ожидаемого риска (особенно для небольших наборов тренировочных данных).

Эмпирический риск - уровень ошибки классификации на тренировочном наборе.

Математическая постановка задачи

Пусть имеется обучающая выборка:

$$(x_1, y_1), \dots, (x_n, y_n), \quad x_i \in \mathbb{R}^n, y_i \in \{-1, 1\}$$

Метод опорных векторов строит классифицирующую функцию F в виде

$$F(x) = \text{sign}(\langle w, x \rangle + b),$$

где $\langle \cdot, \cdot \rangle$ — скалярное произведение, w — нормальный вектор к разделяющей гиперплоскости, b — вспомогательный параметр.

Те объекты, для которых $F(x) = 1$ попадают в один класс, а объекты с $F(x) = -1$ — в другой.

Выбор именно такой функции неслучаен: любая гиперплоскость может быть задана в виде $\langle w, x \rangle + b = 0$ для некоторых w и b .

Математическая постановка задачи

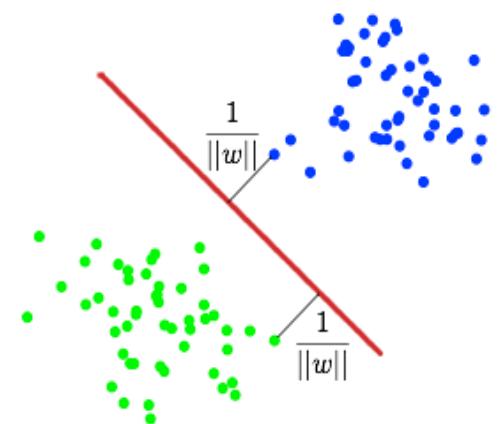
Далее, мы хотим выбрать такие w и b которые максимизируют расстояние до каждого класса.

Можно подсчитать, что данное расстояние равно $\frac{1}{\|w\|}$.

Проблема нахождения максимума эквивалентна проблеме нахождения минимума $\frac{1}{\|w\|}$.

Запишем все это в виде задачи оптимизации:

$$\begin{cases} \arg \min_{w,b} \|w\|^2, \\ y_i(\langle w, x \rangle + b) \geq 1, \quad i = 1, \dots, m. \end{cases}$$



Это является стандартной задачей квадратичного программирования и решается с помощью множителей Лагранжа.

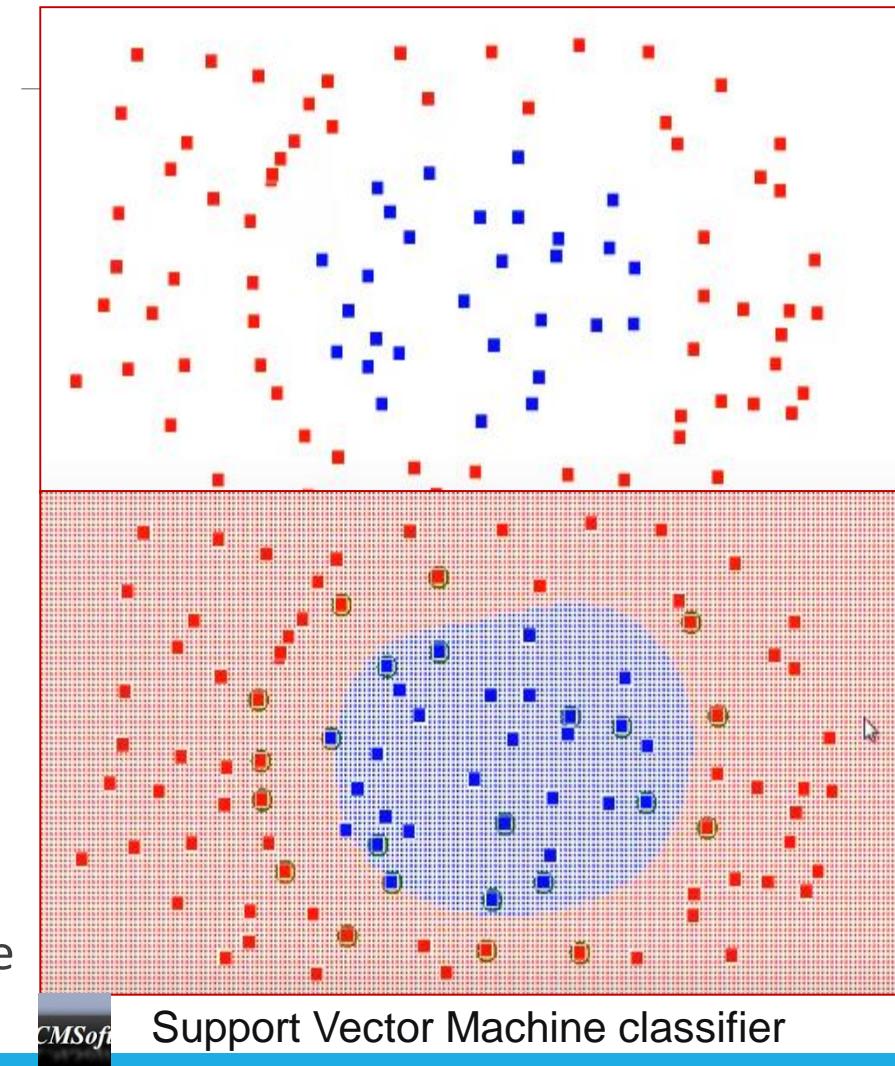
Проблемы классификации в SVM

Одна из проблем – не всегда можно легко найти линейную границу между двумя классами.

В таких случаях один из вариантов - увеличение размерности, т.е. перенос данных из плоскости в трехмерное пространство, где возможно построить такую плоскость, которая идеально разделит множество образцов на два класса.

Опорными векторами в этом случае будут служить объекты из обоих классов, являющиеся экстремальными.

Таким образом, при помощи добавления так называемого **оператора ядра** и дополнительных размерностей, находятся границы между классами в виде гиперплоскостей.



Линейная неразделимость

На практике случаи, когда данные можно разделить гиперплоскостью, или, как еще говорят, *линейно*, довольно редки.

В этом случае поступают так: все элементы обучающей выборки вкладываются в пространство \mathbf{X} более высокой размерности с помощью специального отображения $\varphi; \mathbb{R}^n \rightarrow X$.

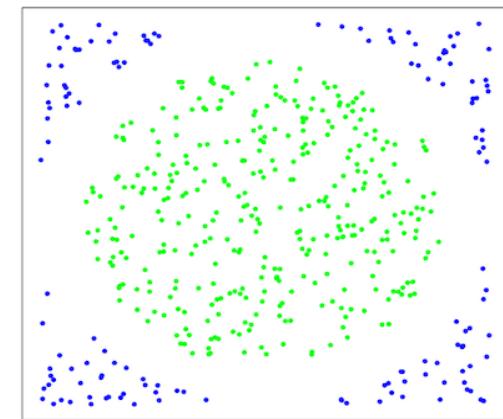
При этом отображение φ выбирается так, чтобы в новом пространстве \mathbf{X} выборка была *линейно* разделима

Классифицирующая функция F принимает вид

$$F(x) = \text{sign}(\langle w, \varphi(x) \rangle + b)$$

Выражение $k(x, x') = \langle \varphi(x), \varphi(x') \rangle$ называется *ядром* классификатора.

С математической точки зрения ядром может служить любая положительно определенная симметричная функция двух переменных. Положительная определенность необходимо для того, чтобы соответствующая функция Лагранжа в задаче оптимизации была ограничена снизу, т.е. задача оптимизации была бы корректно определена.



Чаще всего на практике встречаются следующие ядра:

Полиномиальное:

$$k(x, x') = (\langle x, x' \rangle + \text{const})^d$$

Радиальная базисная функция:

$$k(x, x') = e^{-\gamma \|x - x'\|^2}, \gamma > 0$$

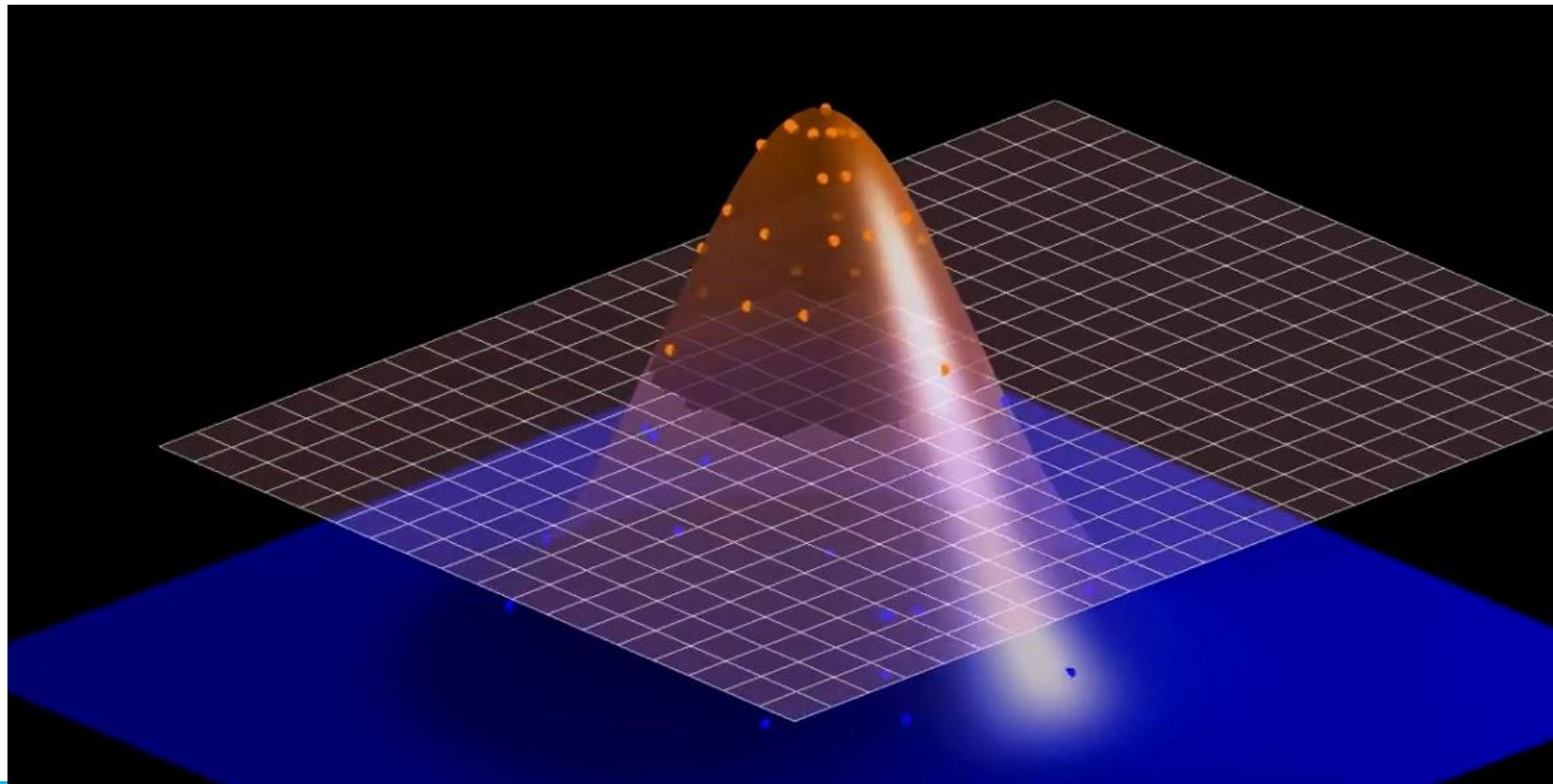
Гауссова радиальная базисная функция:

$$k(x, x') = e^{-\frac{\|x - x'\|^2}{2\sigma^2}}$$

Сигмоид: $k(x, x') = \tanh(\kappa \langle x, x' \rangle + c), \kappa > 0, c > 0$

Сложность построения SVM-модели заключается в том, что чем выше размерность пространства, тем сложнее с ним работать.

Один из вариантов работы с данными высокой размерности - это предварительное применение какого-либо метода понижения размерности данных для выявления наиболее существенных компонент, а затем использование метода опорных векторов.



Достоинства / недостатки SVM

Недостаток метода:

для классификации используется не все множество образцов, а лишь их небольшая часть, которая находится на границах.

Достоинство метода: для классификации методом опорных векторов, в отличие от большинства других методов, достаточно небольшого набора данных.

При правильной работе модели, построенной на тестовом множестве, вполне возможно применение данного метода на реальных данных.

Метод опорных векторов позволяет:

- получить функцию классификации с минимальной верхней оценкой ожидаемого риска (уровня ошибки классификации);
- использовать линейный классификатор для работы с нелинейно разделяемыми данными, сочетая простоту с эффективностью.

Метод «ближайшего соседа»

(«nearest neighbour», «k-nearest neighbour»)

- Относится к классу методов, работа которых основывается на хранении данных в памяти для сравнения с новыми элементами.
- При появлении новой записи для прогнозирования находятся отклонения между этой записью и подобными наборами данных, и наиболее подобная (или ближний сосед) идентифицируется.
- При данном подходе используется термин "*k-ближайший сосед*" ("*k-nearest neighbour*").
- Термин означает, что выбирается k "верхних" (ближайших) соседей для их рассмотрения в качестве множества "ближайших соседей".

Этапы подхода, основанного на прецедентах

Case Based Reasoning, CBR

сбор подробной информации о поставленной задаче;

сопоставление этой информации с деталями прецедентов, хранящихся в базе, для выявления аналогичных случаев;

выбор прецедента, наиболее близкого к текущей проблеме, из базы прецедентов ;

адаптация выбранного решения к текущей проблеме, если это необходимо;

проверка корректности каждого вновь полученного решения;

занесение детальной информации о новом прецеденте в базу прецедентов.

Достоинства / недостатки CBR

Недостатки метода:

- Данный метод не создает каких-либо моделей или правил, обобщающих предыдущий опыт;
- Существует сложность выбора меры "близости" (метрики), также существует высокая зависимость результатов классификации от выбранной метрики;
- При использовании метода возникает необходимость полного перебора обучающей выборки при распознавании, следствие этого - вычислительная трудоемкость;
- Типичные задачи данного метода - это задачи небольшой размерности по количеству классов и переменных.

Преимущества метода:

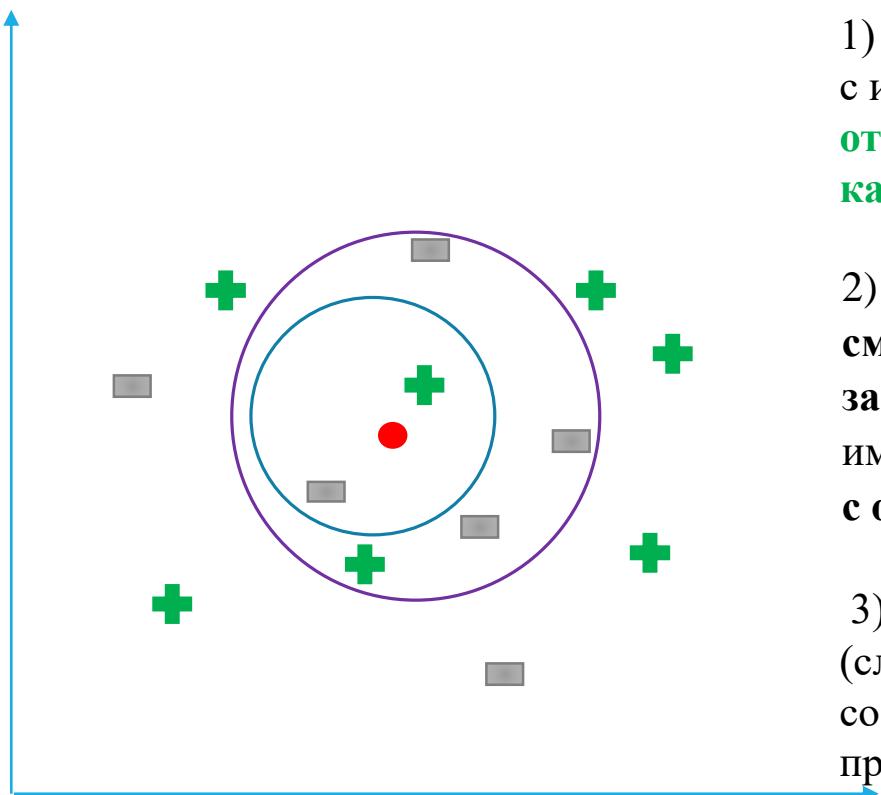
- Простота использования полученных результатов.
- Решения не уникальны для конкретной ситуации, возможно их использование для других случаев.
- Целью поиска является не гарантированно верное решение, а лучшее из возможных;
- С помощью данного метода решаются задачи классификации и регрессии.

Подход, основанный на прецедентах

Case Based Reasoning, CBR

- Вывод, основанный на прецедентах, представляет собой такой метод анализа данных, который делает заключения относительно данной ситуации по результатам поиска аналогий, хранящихся в базе прецедентов.
- Данный метод по сути относится к категории "**обучение без учителя**", (является "самообучающейся" технологией), благодаря чему рабочие характеристики каждой базы прецедентов с течением времени и накоплением примеров улучшаются.
- Разработка баз прецедентов по конкретной предметной области происходит на естественном для человека языке (может быть выполнена наиболее опытными сотрудниками компании - экспертами или аналитиками, работающими в данной предметной области).

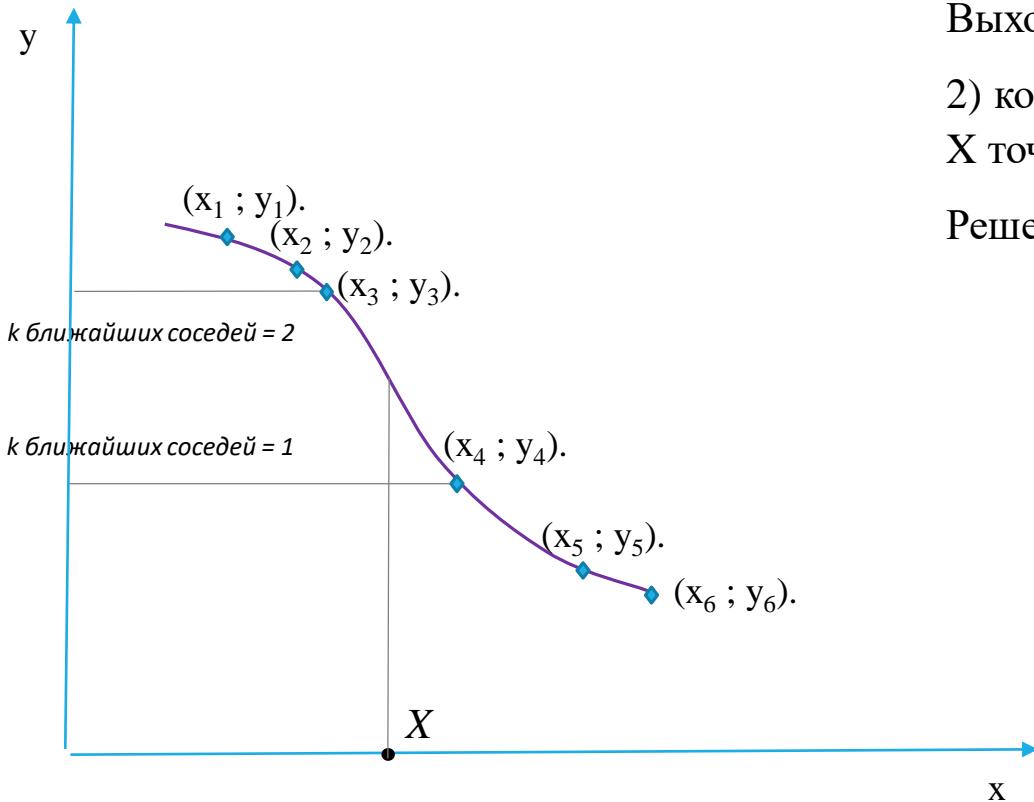
Решение задачи классификации новых объектов



- 1) Результат работы метода k -ближайших соседей с использованием одного ближайшего соседа:
отклик точки запроса будет классифицирован как знак плюс
- 2) Метод k -ближайших соседей (в случае 2) не сможет классифицировать отклик точки запроса, поскольку вторая ближайшая точка имеет знак минус и оба знака равнозначны (**победа с одинаковым количеством голосов**);
- 3) Результат работы метода k -ближайших соседей (случай 5 соседей): 2 точки со знаком "+" и 3 точки со знаком "-", алгоритм k -ближайших соседей присвоит знак "-" отклику точки запроса

Классификация объектов множества при разном значении параметра k

Решение задачи прогнозирования



1) метод k -ближайших соседей при $k = 1$, ищем набор примеров и выделяем из их числа ближайший к точке запроса X .

Выход Y равен y_4 ($Y = y_4$).

2) когда $k = 2$, выделяем уже две ближайшие к X точки (точки y_3 и y_4 соответственно).

Решение для Y в виде $Y = (y_3 + y_4)/2$.

Решение задачи прогнозирования осуществляется путем переноса описанных выше действий на использование произвольного числа ближайших соседей таким образом, что **выход Y точки запроса X вычисляется как среднеарифметическое значение выходов k -ближайших соседей точки запроса.**

Метод «ближайшего соседа» для задачи прогнозирования

- Независимые и зависимые переменные набора данных могут быть как непрерывными, так и категориальными. Для непрерывных зависимых переменных задача рассматривается как задача прогнозирования, для дискретных переменных - как задача классификации.
- Предсказание в задаче прогнозирования получается усреднением выходов k-ближайших соседей, а решение задачи классификации основано на принципе "*по большинству голосов*".
- Критическим моментом в использовании метода k-ближайших соседей является выбор параметра k. Он один из наиболее важных факторов, определяющих качество прогнозной либо классификационной модели.
- *Должно быть выбрано оптимальное значение параметра k* (это значение должно быть настолько большим, чтобы свести к минимуму вероятность неверной классификации, и одновременно, достаточно малым, чтобы k соседей были расположены достаточно близко к точке запроса).

Рассматриваем k как сглаживающий параметр, для которого должен быть найден компромисс между силой размаха (разброса) модели и ее смещенностью.

Метод «ближайшего соседа» для задачи прогнозирования

Оценка параметра k методом кросс-проверки

Один из вариантов оценки параметра k - проведение кросс-проверки (Bishop, 1995).

Кросс-проверка - известный метод получения оценок неизвестных параметров модели. Основная идея метода - разделение выборки данных на v "складки". В "складки" здесь случайным образом выделенные изолированные подвыборки.

По фиксированному значению k строится модель k -ближайших соседей для получения предсказаний на v -м сегменте (остальные сегменты при этом используются как примеры) и оценивается ошибка классификации.

Для регрессионных задач наиболее часто в качестве оценки ошибки выступает сумма квадратов, а для классификационных задач удобней рассматривать точность (процент корректно классифицированных наблюдений).

Второй вариант выбора значения параметра k - самостоятельно задать его значение. Однако этот способ следует использовать, если имеются обоснованные предположения относительно возможного значения параметра, например, предыдущие исследования сходных наборов данных.

Метод k -ближайших соседей показывает достаточно неплохие результаты в самых разнообразных задачах.

Инструменты Data Mining, реализующих метод k-ближайших соседей и CBR-метод:

CBR Express и Case Point (Inference Corp.),

Apriori (Answer Systems), DP Umbrella (VYCOR Corp.),
KATE tools (Acknosoft, Франция),

Pattern Recognition Workbench (Unica, США),

а также некоторые статистические пакеты, например,
Statistica и др.

Литература

- B. Scholkopf, G. Ratsch, K. Muller, K. Tsuda, S. Mika An Introduction to Kernel-Based Learning Algorithms / IEEE Neural Networks, 12(2):181-201, May 2001
- Chickering D, Geiger D., Heckerman D. Learning Bayesian networks: The combination of knowledge and statistical data / Machine Learning. 1995. 20. P. 197-243
- Heckerman D Bayesian Networks for Data Mining / Data Mining and Knowledge Discovery. 1997. № 1. P. 79-119
- etc, Friedman N., Geiger D., Goldszmidt M. Bayesian Network Classifiers / Machine Learning. 1997. 29. P. 131-165
- Brand E., Gerritsen R / Naive-Bayes and Nearest Neighbor DBMS. 1998. № 7