Обзор методов машинного обучения

Курс «Интеллектуальные информационные системы»

Кафедра управления и информатики НИУ «МЭИ»

Осень 2017 г.

Метод потенциальных функций

Общая идея метода иллюстрируется на примере электростатического взаимодействия элементарных частиц.

Классифицируемый документ относится к классу, чей наведенный совокупный потенциал Φ_k выше

$$\Phi_k(\vec{X}_{N+1}) = \frac{\sum_{j=1}^{N_k} \varphi(\rho(\vec{X}_{N+1}, \vec{X}_j))}{N_k}$$

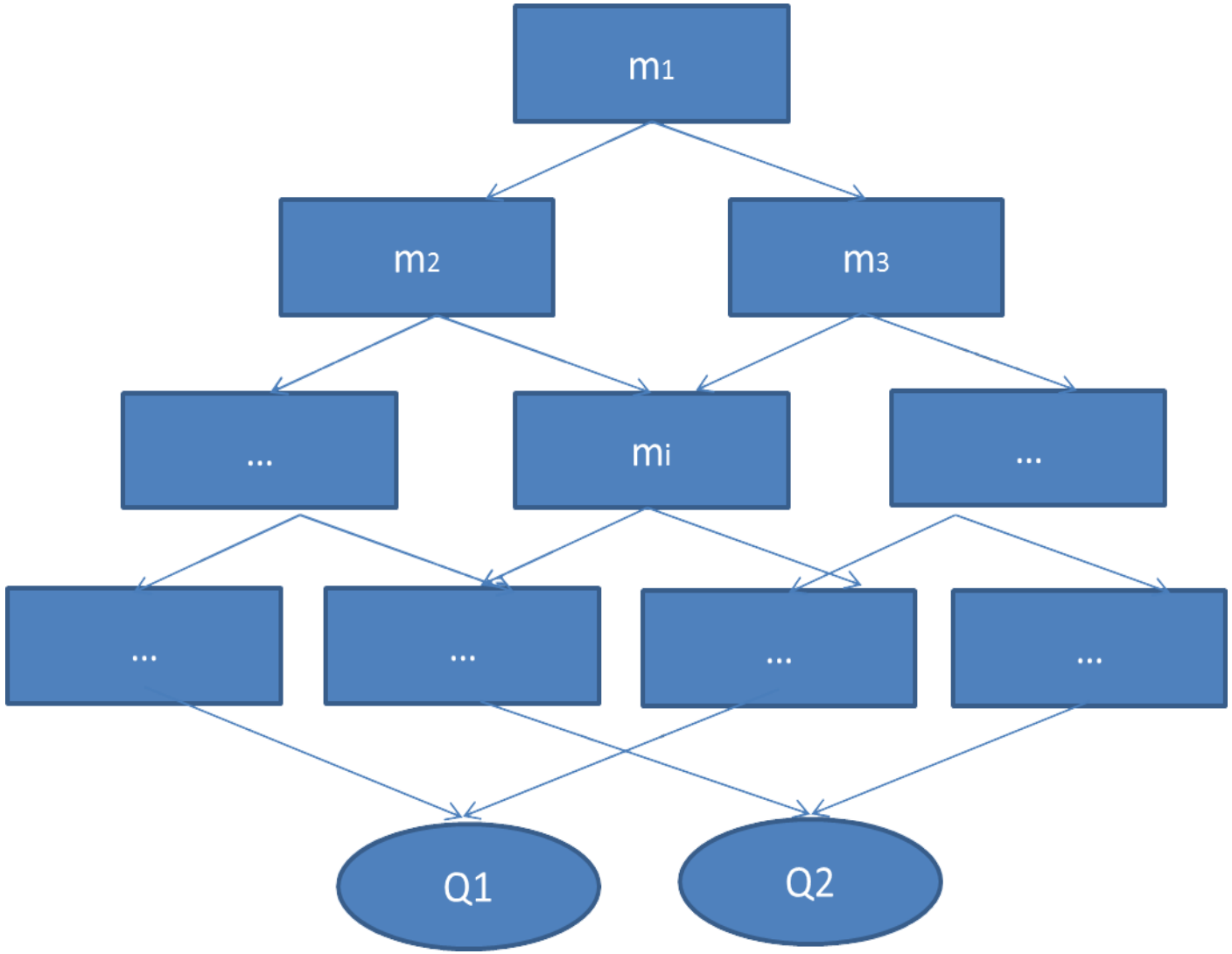
$\varphi(\rho)$ – некоторая известная положительная функция от метрики расстояния

$$\varphi(\rho) = \frac{1}{\rho + \alpha} \qquad \varphi(\rho) = e^{-\alpha\rho^{\beta}} \qquad \varphi(\rho) = (1 + \alpha\rho^{\beta})^{-1}$$

Метод деревьев решений

Средство поддержки принятия решений, использующееся в статистике и анализе данных для прогнозных моделей.

В методе деревьев решений проводится последовательное разделение множества документов на основе значений выбранного признака, в результате чего строится дерево, содержащее нетерминальные узлы (узлы проверок), в которых происходит разбиение по выбранному атрибуту, и терминальные узлы (узлы ответа), в которых должны находиться элементы одного класса.



Метод деревьев решений. Критерий прироста информации

Для выбора наиболее информативного признака, по которому проводится разбиение, в методе деревьев решений чаще всего используется *теоретико-информационный (энтропийный) подход*.

Хотим найти такой признак $x^{(S)}$, при разбиении по которому один из классов имел наибольшую вероятность появления. Это возможно, если величина прироста информации *Gain* будет достигать своего максимума.

$$Gain(x^{(s)},T) = I(T) - I(x^{(s)},T)$$

$$I(T) = \sum_{k=1}^{K} P_k \log_2 \frac{1}{P_k} = -\sum_{k=1}^{K} \frac{N_k}{N} \log_2 \frac{N_k}{N}$$

- среднее количество информации (энтропия), необходимое для определения класса примера из обучающей выборки *T*

$$I(x^{(s)},T) = \sum_{i=1}^{S} \frac{N_s}{N} I(T_s) = \sum_{s=1}^{S} \frac{N_s}{N} \left(-\sum_{k=1}^{K} \frac{N_{ks}}{N_s} \log_2 \frac{N_{ks}}{N_s} \right)$$

- среднее количество информации, необходимое для идентификации класса примера в каждом подмножестве после разбиения по признаку $x^{(S)}$

Метод деревьев решений. Меры неоднородности

Еще один подход к выявлению признака, по которому стоит проводить разбиение – использовать меры неоднородности ф. Здесь вектор **p** состоит из **m** вероятностей меток встречающихся в некотором подмножестве обучающего множества

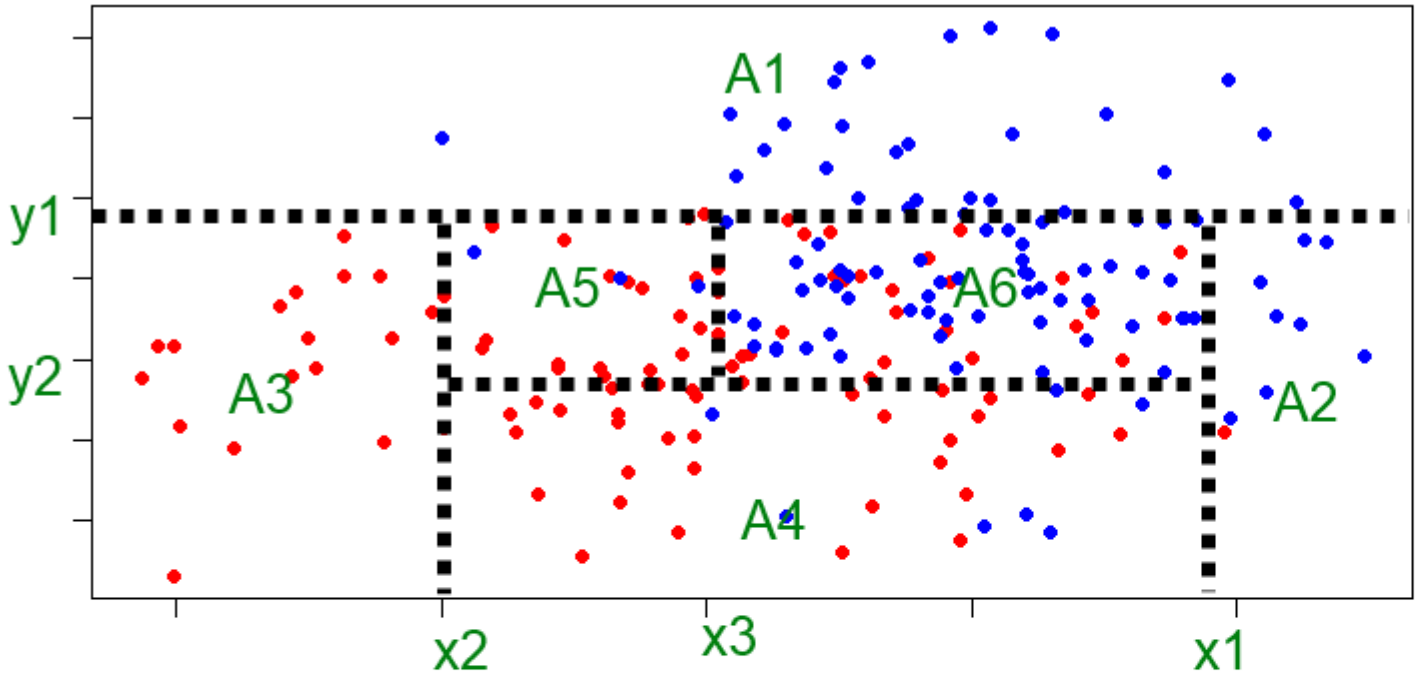
$$\phi(\vec{p}) = 1 - \max(\vec{p})$$ Наиболее часто встречаемый класс

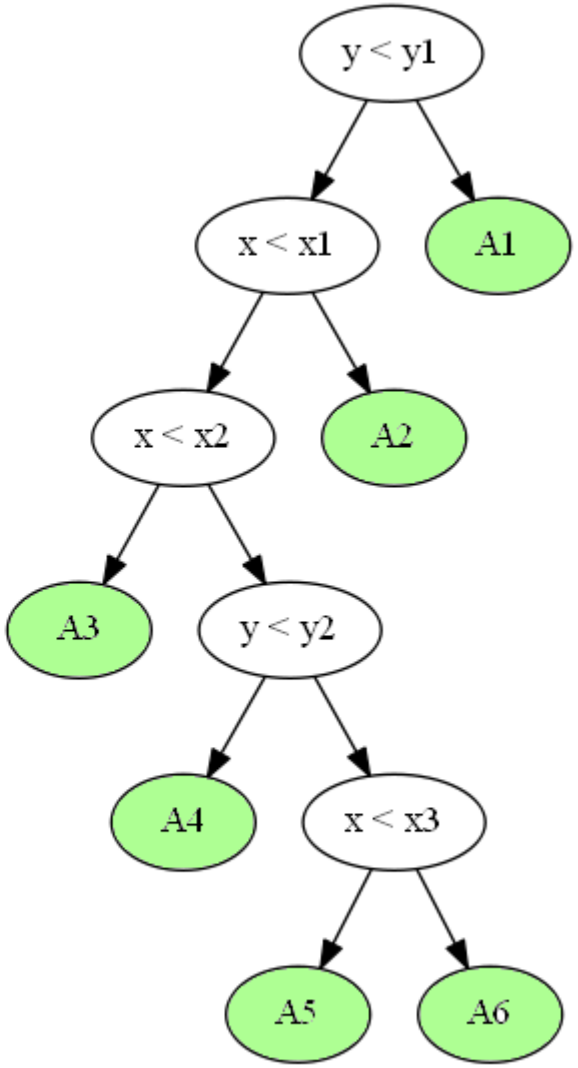
$$\phi(\vec{p}) = \sum_{i=1}^{m} p_i(1 - p_i)$$ Индекс (коэффициент) Джини (Gini index)

$$\phi(\vec{p}) = -\sum_{i=1}^{m} p_i \log(p_i)$$ Перекрестная энтропия

На каждой итерации для входного подмножества обучающего множества строится такое разбиение пространства гиперплоскостью (ортогональной одной их осей координат), которое минимизировало бы среднюю меру неоднородности двух полученных подмножеств. Данная процедура выполняется рекурсивно для каждого полученного подмножества до тех пор, пока не будут достигнуты критерии остановки.

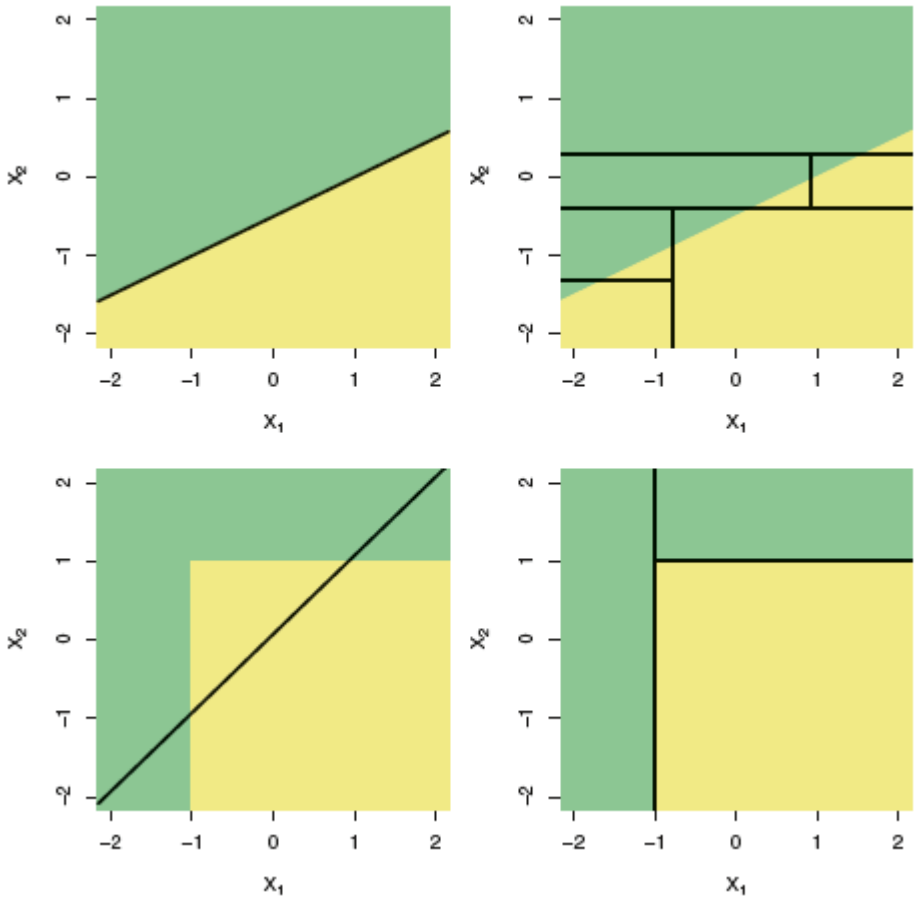
Метод деревьев решений. Пример разбиения





Метод деревьев решений. Область использования

Сравнение линейных алгоритмов и алгоритмов, основанных на деревьях решений:



Недостатки:

- Алгоритмы «локальны», не могут обеспечить оптимальность всего дерева в целом
- Свойственна проблема «переобученности»

Достоинства

- Прост в понимании и интерпретации
- Не требует предварительной обработки данных
- Метод хорошо работает даже в том случае, если были нарушены первоначальные предположения, включенные в модель.

Случайный лес (Random Forest)

Пусть обучающая выборка состоит из N примеров, размерность пространства признаков равна M, и задан параметр m

Все деревья комитета строятся независимо друг от друга по следующей процедуре:

1. Сгенерируем случайную подвыборку **с повторением** размером *N* из обучающей выборки. (Таким образом, некоторые примеры попадут в неё несколько раз, а в среднем $N(1 - 1/N)^N$, т.е. примерно *N/e* примеров не войдут в неё вообще)
2. Построим решающее дерево, классифицирующее примеры данной подвыборки, причём в ходе создания очередного узла дерева будем выбирать признак, на основе которого производится разбиение, не из всех M признаков, а лишь из m случайно выбранных.
3. Проводится построение дерева

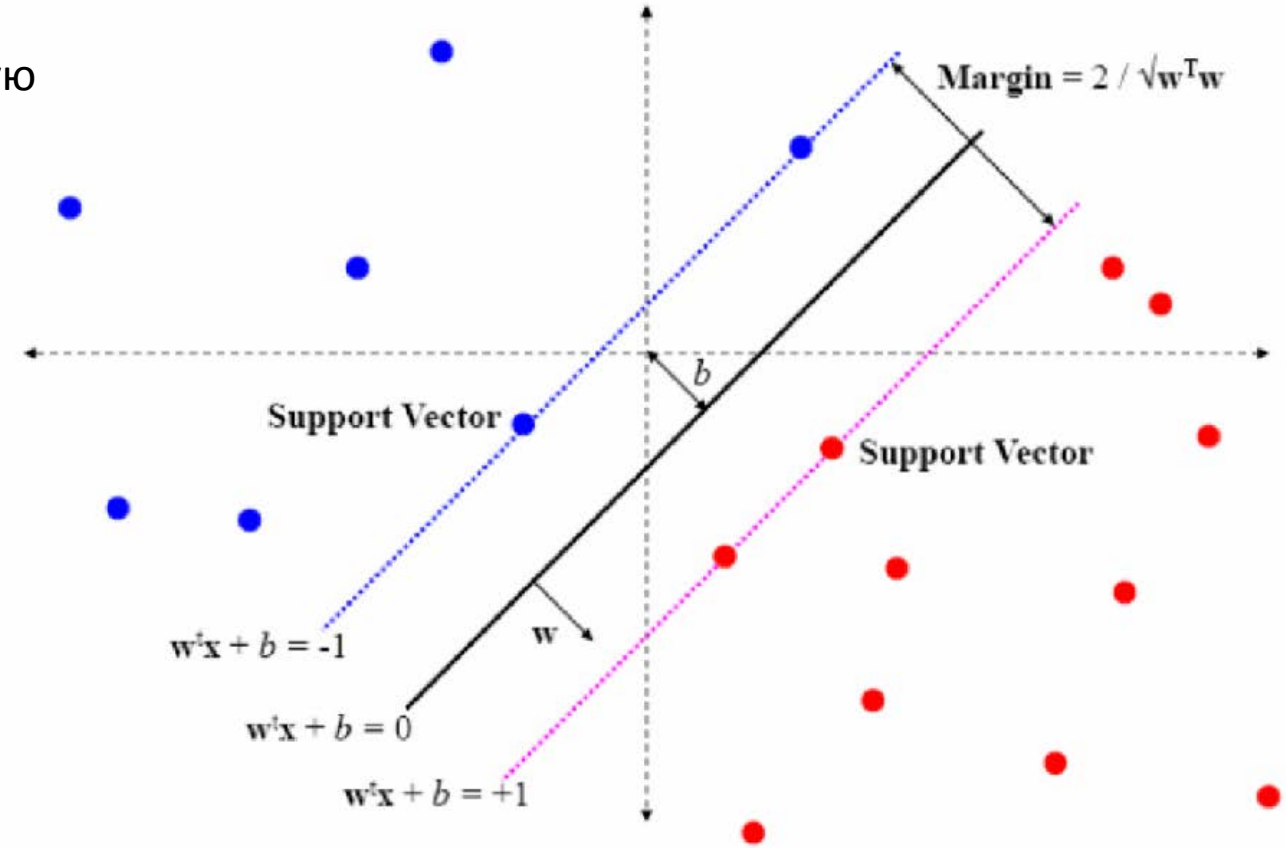
Классификация объектов проводится путём голосования: каждое дерево комитета относит классифицируемый объект к одному из классов, и побеждает класс, за который проголосовало наибольшее число деревьев.

Метод опорных векторов (SVM, Support Vector Machine)

Алгоритм предложен в 1963 году Владимиром Вапником и Алексеем Червоненкисом. Основная идея метода — перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей классы. Разделяющей гиперплоскостью будет гиперплоскость, максимизирующая расстояние до двух параллельных гиперплоскостей.

Метод опорных векторов строит классифицирующую функцию F в виде: $F(\mathbf{x}) = \operatorname{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + b)$

где w — нормальный вектор к разделяющей гиперплоскости, b — вспомогательный параметр

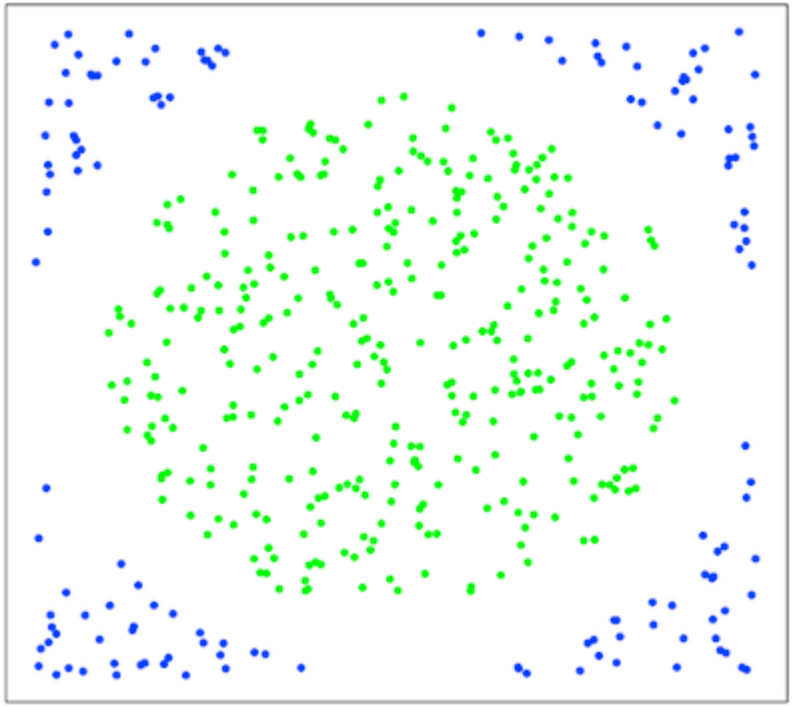


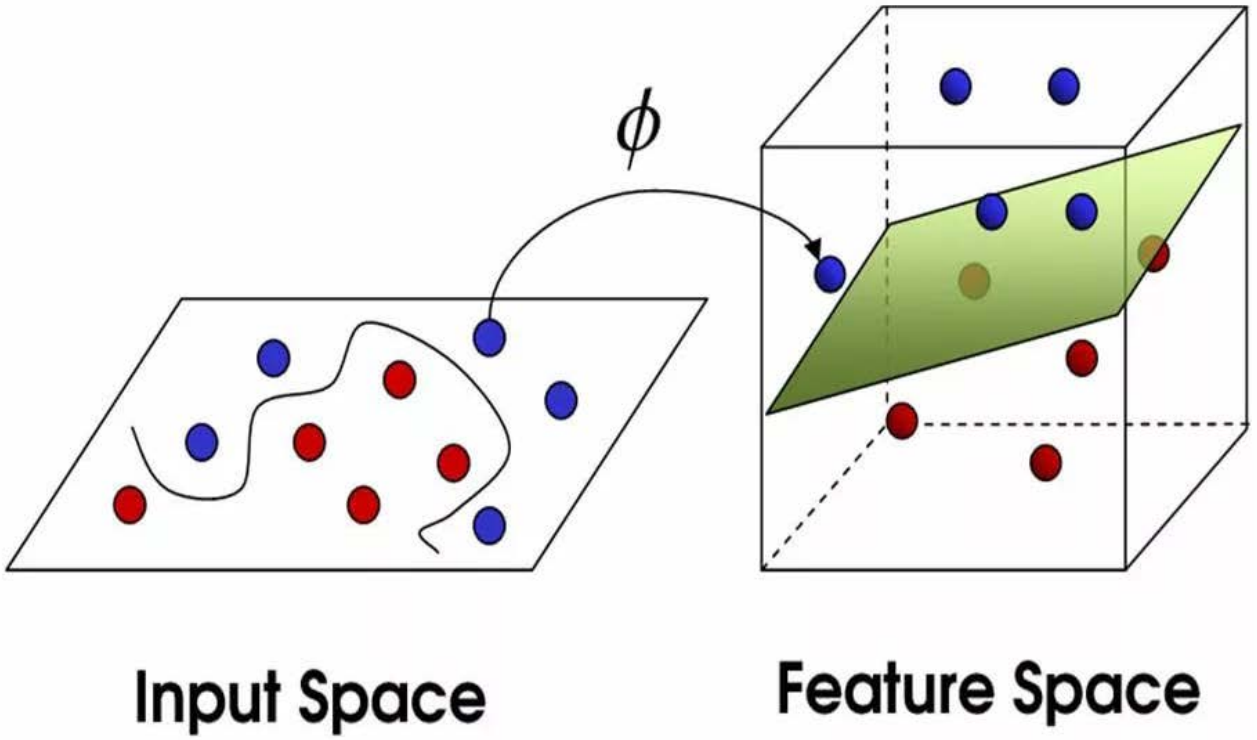
Далее выбираются такие w и b, которые

максимизируют расстояние $\frac{1}{\|w\|}$ до каждого класса

Метод опорных векторов. Линейная неразделимость

Если данные линейно неразделимы, то все элементы обучающей выборки вкладываются в пространство **X** более высокой размерности с помощью специального отображения[1] $\phi : R^n \to X$





[1]**К. В. Воронцов.** *Лекции по методу опорных векторов.* http://www.ccas.ru/voron/download/SVM.pdf

Логистическая регрессия

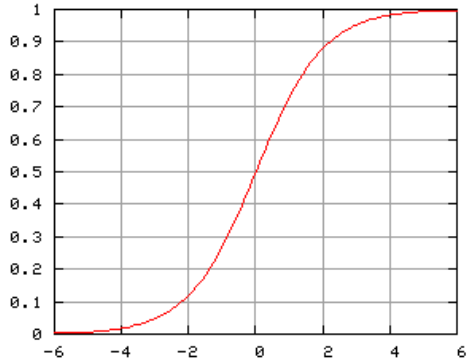
Логистическая регрессия (Logistic regression) — метод построения линейного классификатора, позволяющий оценивать апостериорные вероятности принадлежности объектов классам.

Алгоритм классификации: $$a(x,w) = \operatorname{sign}\left(\sum_{j=1}^{n} w_j f_j(x) - w_0\right) = \operatorname{sign}\langle x,w\rangle,$$

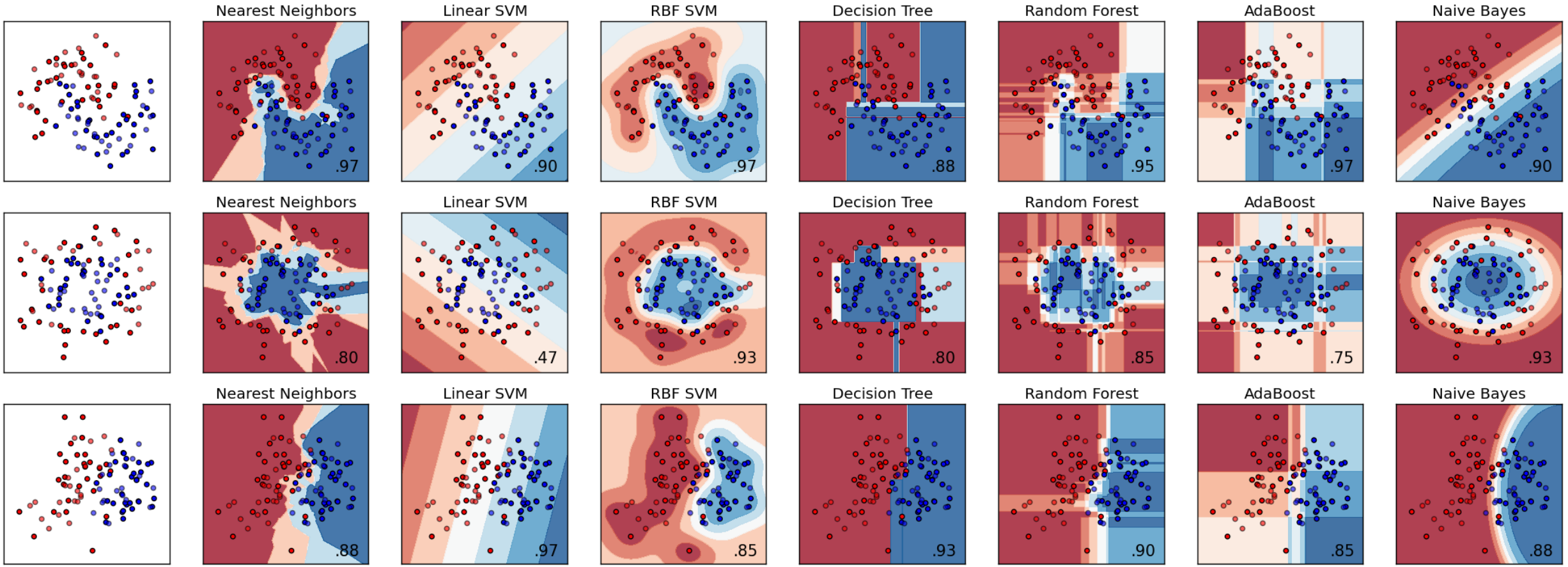
Задача обучения линейного классификатора заключается в том, чтобы по обучающей выборке настроить вектор весов *w*. В логистической регрессии для этого решается задача минимизации эмпирического риска с функцией потерь специального вида:

$$Q(w) = \sum_{i=1}^{m} \ln\left(1+\exp(-y_i\langle x_i,w\rangle)\right) \to \min_w.$$

После того, как решение *w* найдено, становится возможным оценить апостериорную вероятность принадлежности объекта классу с помощью логистической функции: $$f(x) = \frac{1}{1+e^{x}} \quad \Rightarrow \quad P\{y \mid x\} = \frac{1}{1+e^{y\langle x,w\rangle}}$$



Сравнение методов на разных типах выборок



Тематическое моделирование

Тематическая модель (topic model) — модель коллекции текстовых документов, которая определяет, к каким темам относится каждый документ коллекции. Алгоритм построения тематической модели получает на входе коллекцию текстовых документов. На выходе для каждого документа выдаётся числовой вектор, составленный из оценок степени принадлежности данного документа каждой из тем. Размерность этого вектора, равная числу тем, может либо задаваться на входе, либо определяться моделью автоматически.

Предполагается, что каждый документ может относиться к одной или нескольким темам. Темы отличаются друг от друга различной частотой употребления слов. Требуется найти эти темы, то есть определить

- число тем;
- распределения частот слов, характерное для каждой темы;
- тематику каждого документа — в какой степени он относится к каждой из тем.

Данная задача может рассматриваться как задача одновременной кластеризации *документов* и *слов* по одному и тому же множеству кластеров, называемых *темами.*

Литература: http://www.machinelearning.ru/wiki/images/e/e6/Voron-ML-TopicModeling-slides.pdf

Тематическое моделирование

- Тема – семантически однородный кластер
- Тема – специальная терминология предметной области
- Тема – набор терминов, совместно встречающихся в документах

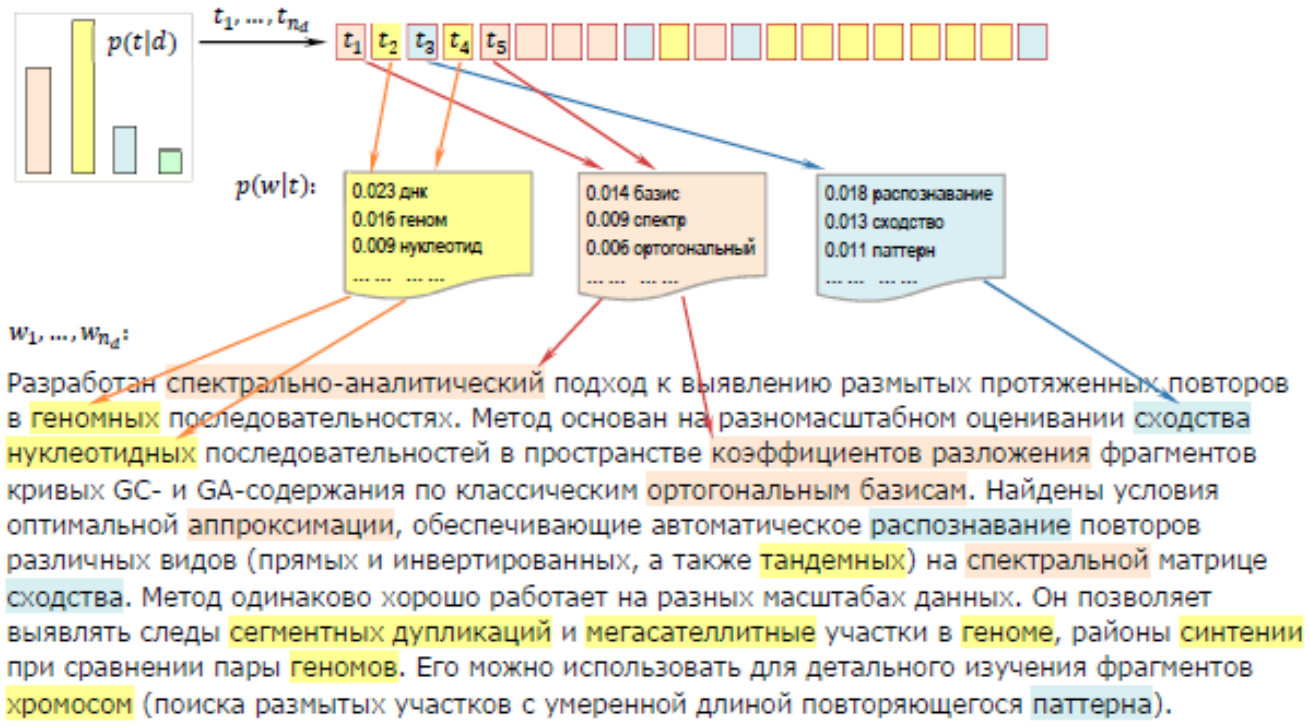
Более формально:
Тема – условное распределение на множестве терминов
$p(w|t)$ – вероятность термина w в теме t
Тематический профиль документа - условное распределение
$P(t|d)$ – вероятность темы t в документе d

Когда автор писал термин w в документе d, он думал о теме t, и мы хотели бы узнать, о какой именно.

Тематическое моделирование

Вероятностная тематическая модель коллекции документов D описывает появление терминов w в документе d:

$$p(w|d) = \sum p(w|t)p(t|d)$$



Построение модели с помощью:

- Латентное размещение Дирихле (LDA)
- Вероятностный латентно-семантический анализ (pLSA)

Литература: http://www.machinelearning.ru/wiki/index.php?title=Тематическая_модель