**Общая постановка задачи обучения***.* Модель обучения по эмпирическим данным предполагает наличие:

* генератора случайных входных объектов ***x*** – элементов некоторого пространства *X∈Rn*, появляющихся независимо согласно фиксированному, но неизвестному распределению вероятностей *P*(***x***);
* учителя, определяющего выход *y* – элемента некоторого пространства *Y*, для любого входящего ***x***, согласно условному распределению *P*(*y*/***x***), также фиксированному и неизвестному;
* класса функций *F*={*f*(***x***)}

(в задаче бинарной классификации

*F* = { *f* (***x***): *Rn* →{±1} }).

Задача обучения состоит в выборе из заданного множества функций одной функции, которая предсказывает ответ учителя наилучшим образом. Этот выбор должен быть основан на тренировочной последовательности (ТП) конечного объема (*l*), т.е. независимых, одинаково распределенных согласно закону *P*(***x***,*y*) *= P*(***x***)*×P*(*y*/***x***) наблюдениях (***x****1*, *y1*), … , (***x****l*, *yl*).

Наилучшая функция *f*, которую можно выбрать, – функция, минимизирующая ожидаемый риск

,

где *L* − содержательно обоснованная функция потерь. В частности,

.

Поскольку распределение *P*(***x***,*y*) неизвестно, можно, руководствуясь индукционным принципом минимизации эмпирического риска, заменить среднее по мере *P*(***x***,*y*) средним по тренировочным данным:



Однако для конечных выборок этот принцип оказывается несостоятельным, поскольку может приводить к переобучению (overfitting). Об эффекте переобучения говорят, если качество решающей функции вне ТП, на вновь поступающих образцах, оказывается существенно хуже качества, достигнутого на ТП.

Один из путей преодоления явления переобучения состоит в сужении класса аппроксимирующих функций до класса со сложностью, подходящей для имеющейся тренировочной последовательности.

Статистическая теория обучения (Vapnik V.N*.* The Nature of Statistical Learning Theory. Springer-Verlag, 2000) предлагает **новый индукционный принцип для обучения по конечным выборкам – структурную минимизацию риска, сущность которого состоит в минимизации верхней границы ожидаемого риска, что обеспечивает согласованность между качеством обучения и сложностью класса аппроксимирующих функций.** При условии, что эмпирический риск с ростом объема ТП стремится к ожидаемому риску, верхняя граница ожидаемого риска, основанная на некоторой мере сложности функций *h*, имеет вид:

 (1)

и гарантирована с вероятностью 1–η для любого η∈(0,1) и любой функции *f* из заданногокласса аппроксимирующих функций {*f(****x****)*} с мерой сложности *h* для *l* > *h*. Второе слагаемое в (1) контролирует сложность класса аппроксимирующих функций и не зависит от ТП.

Принцип структурной минимизации риска предлагает определенный способ, позволяющий контролировать сложность класса функций, из которого выбирается решение. На заданном множестве функций {*f*(***x***)} вводится вложенная структура согласно некоторой мере сложности подклассов. На каждом элементе структуры выбирается функция*,* минимизирующая эмпирический риск. Затем из отобранных функций выбирается функция, доставляющая минимум верхней границе ожидаемого риска (1).

**Принцип структурной минимизации риска оснащает SV-машины способностью к обобщению, которая и является целью статистического обучения.**

**Линейный SVM-классификатор**

**Линейно разделимые данные.** Рассмотрим на примере задачи бинарной классификации линейно разделимых данных основные идеи и преимущества SVM-обучения.Пусть (***x****1*,*y1*), … , (***x****l*, *yl*)∈*Rn*×*R* − тренировочная последовательность. В этом случае:

* эмпирический риск должен быть равным нулю;
* выход *y* может принимать только два значения: −1 и +1 (метка класса);
* заданный класс аппроксимирующих функций – множество гиперплоскостей *f*(***x***)=0, где *f*(***x***)=(***w***, ***x***)+*b*, (***w***, ***x***) – скалярное произведение ***x*** и ***w*** ∈*Rn* , *b*∈*R*;
* в качестве функции потерь естественно взять δ(–*yf*(*x*)), где δ(⋅) – функция Хевисайда.

Выберем масштаб для параметров разделяющей гиперплоскости (ГП) так, чтобы для векторов ТП выполнялось условие

.

Тогда расстояние между двумя ближайшими к такой канонической разделяющей ГП векторами из разных классов (так называемая margin, т.е. ширина полосы) будет равно 2/||***w***||, где  – евклидова норма. Гиперплоскости вида (***w***, ***x***)+*b* = ±1 называются поддерживающими. На рис. 1 для *n*=2 приведены каноническая разделяющая ГП, соответствующие поддерживающие гиперплоскости и указана ширина полосы. Поскольку целью обучения является машина, обладающая, насколько возможно, лучшей способностью к обобщению, оптимальным решением задачи классификации должна быть каноническая разделяющая ГП, имеющая максимальную margin.

Введем на множестве канонических разделяющих ГП вложенную структуру с использованием такой меры сложности классов гиперплоскостей как размерность Вапника-Червоненкиса (*h*), а затем, в процессе обучения, выберем подкласс с *h*, подходящей для данной тренировочной последовательности, при которой верхняя граница ожидаемого риска (1) будет минимальной. Вапник показал, что *h* для класса гиперплоскостей ограничена сверху величиной, связанной с шириной полосы: если ограничить снизу ширину полосы величиной 2/*A*, то ||***w***||≤*A* и *h*≤*A2R2+*1, где *R* – радиус наименьшей гиперсферы, охватывающей множество векторов ТП. Поэтому малое значение ||***w***|| приводит к малому *h*, и минимизация ||***w***|| представляет собой реализацию принципа структурной минимизации риска. Таким образом, минимум нормы веса ||***w***|| канонической разделяющей ГП максимизирует величину margin и минимизирует Ω(*h*) одновременно.

Найдем



которые являются условиями классификации без ошибок. Составим функцию Лагранжа, соответствующуюэтой задаче:

 (2)

Здесь α*i*≥0 − множители Лагранжа. Сформулируем прямую задачу в терминах целевой функции *Lp*: найти минимум целевой функции *Lp* по ***w*** и *b* требуя, чтобы градиент лагранжиана по всем {α*i*} равнялся нулю, и все α*i*≥0***.*** Это задача выпуклого квадратичного программирования. Вместо нее можно решать эквивалентную ей двойственную задачу: найти максимум целевой функции*Lp*повсем {α*i*}, требуя, чтобы градиент лагранжианапо***w*** и*b*равнялся нулю, и всеα*i*≥ 0***.***

Вычислив градиент лагранжиана  по ***w*** и *b* и приравняв его нулю, получим:

 (3)

 (4)

Подставив полученные ограничения типа равенств в лагранжиан (2), представим окончательную формулировку двойственной задачи:

 (5)

Таким образом, решение задачи SVM-обучения эквивалентно нахождению решения задачи квадратичного программирования (5), удовлетворяющего условиям Каруша-Куна-Таккера (ККТ):

 (6)

 (7)

 (8)

 (9)

 (10)

Условия ККТ для этой задачи являются необходимыми и достаточными. Использование условий ККТ влечет за собой ряд очевидных преимуществ:

* ***w*** находят, согласно (3), с использованием только ТП. Такой вид вектора весов называют SV-разложением;
* при α*i*>0 из уравнения (10) легко найти *b*;
* правильно классифицированным векторам ТП, лежащим вне полосы, соответствуют α*i*= 0;
* правильно классифицированным векторам ТП, лежащим наоптимальных поддерживающих гиперплоскостях, соответствуют α*i*>0, эти векторы и будут опорными, определяющими решение задачи.

Опорные векторы − наиболее информативные точки тренировочной последовательности: если убрать из нее все остальные точки и повторить процесс обучения, решение останется тем же самым.

Окончательное решение задачи бинарной классификации линейно разделимых данных (hard margin SVM) представляет собой оптимальную гиперплоскость, имеющую вид:

 (11)

Для тестирования новых образцов используют линейную машину опорных векторов:

 (12)

Количество свободных параметров равно количеству опорных векторов (число которых значительно меньше объема *l* ТП) и не зависит от размерности *n* входных векторов.

**Линейно неразделимые данные.** В случае линейно неразделимых данных выберем каноническую гиперплоскость, которая разделяет тренировочные точки настолько правильно, насколько это возможно, следуя идее максимальной ширины полосы, т.е. построим soft margin SVM.

Для ослабления жестких ограничений вводят неотрицательные переменные ξ*i*, *i*=1,…,*l*. Теперь в процессе обучения, наряду с ||***w***||, минимизируют и верхнюю границу эмпирического риска:



где параметр регуляризации *C* > 0 контролирует ширину полосы, значение параметра задается. Чем больше *C*, тем уже ширина полосы. *С* − штрафной параметр, его большее значение соответствует назначению большего штрафа ошибкам классификации тренировочных образцов. Соответствующий лагранжиан имеет вид:

, (13)

где, μ*i*− множитель Лагранжа для ограничений на переменные ξ*i*, *i*=1, …, *l*.

Вычислив градиент лагранжиана *Lp* по ***w***, *b* и **ξ**, приравняв его нулю, подставив полученные ограничения типа равенств в лагранжиан (13), представим окончательную формулировку двойственной задачи для линейно неразделимого случая:



В двойственной задаче не фигурируют ни переменные ξ*i*, *i*=1, …, *l*, ни множители μ*i*, а параметр *С*  появляется только в ограничениях. Выпишем условия оптимальности (условия ККТ) для этой задачи:

 (14)

 (15)

 (16)

 (17)

 (18)

 (19)

 (20)

  (21)

 (22)

Из соотношений (21) и (22) определяем сдвиг *b,* заметив, что ξ*i* = 0, если α*i* <*C* (это следует из (16) и (22)).

Условия (14)–(22) для оптимальных множителей Лагранжа позволяют разбить множество векторов ТП на три группы:

* правильно классифицированные векторы ***x****i*: α*i*=0, ξ*i*=0
* опорные векторы ***x****i*: 0<α*i*<C, ξ*i*=0 (USVs)
* ошибочно классифицированные векторы ***x****i*: α*i*=C, ξ*i*>0 (BSVs)

(или *связанные* опорные векторы)

Из условий ККТ, следует, что решающее правило для линейно неразделимых данных имеет тот же вид, что и в линейно разделимом случае:

,

и по-прежнему является разреженным по **α**.