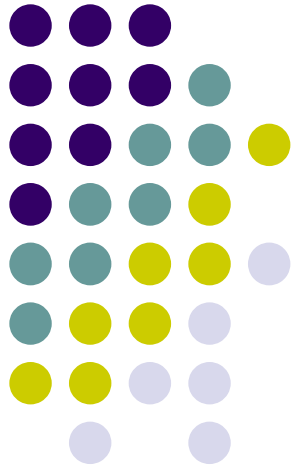
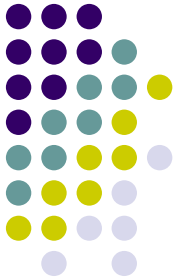
Статистическая природа обучения

Корлякова М.О.

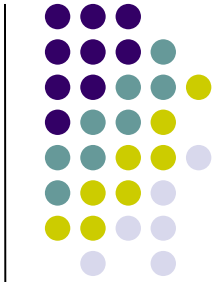
2018



Оценка классификатора



VC - измерение



- ограничение на скорость сходимости
- измерение Вапника-Червоненкиса (1971)

- мера емкости (вычислительной мощности) семейства функций классификации, реализованных обучаемыми машинами.

Определение VC

- VC-измерением классификатора F называют мощность наибольшего множества L, разбиением которого является F.

- VC-измерение комбинаторно

двоичная классификация

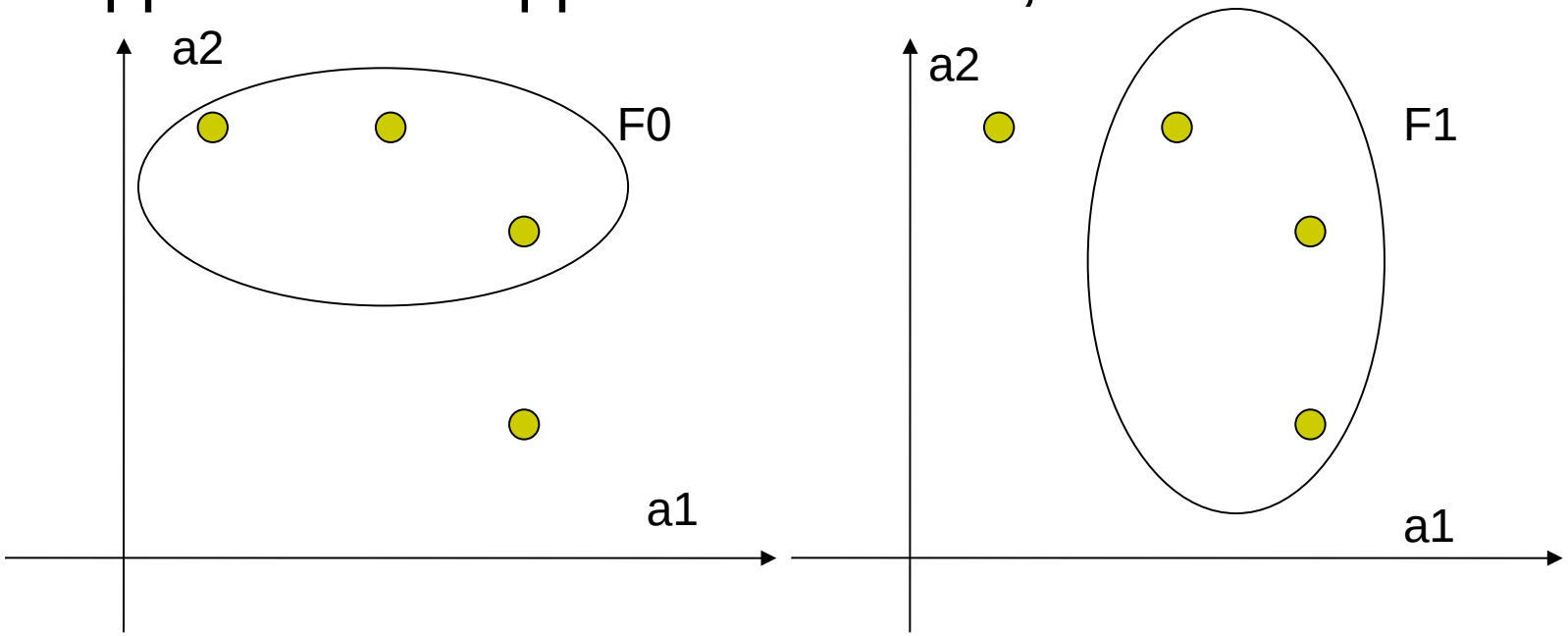
- d={0,1}

- правило F(**x**,w) – i-я дихотомия
- множество дихотомий, реализованных обучаемой машиной

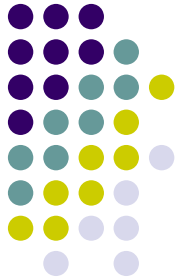
  $F = \{F(\mathbf{x},w) : w \in W, F: \Re^m W \to \{0,1\}\}$

- L = {xi∈X: i=1.N} – точки m-мерного пространства

- дихотомия делит L на L1, L0

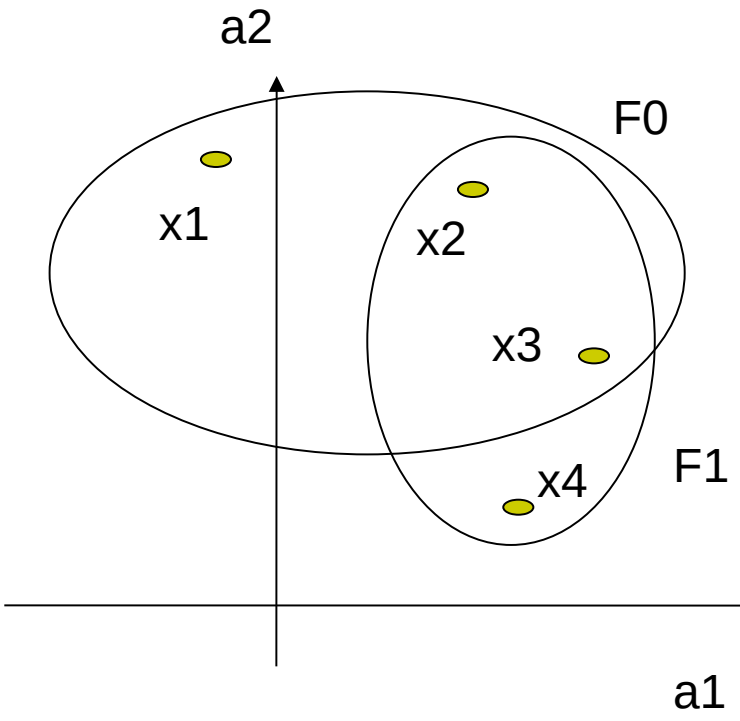


F(x,w)={ 0 для x∈L0 , 1 для x∈L1}



- $\Delta_F(L)$ - число различных дихотомий
- $\Delta_F(\ell)$ - максимум для $\Delta_F(L)$ всех L, где $|L|=\ell$
- $\Delta_F(\ell)$ - функция роста.

Пример



D0={G0={x1,x2,x3}}, G1={x4}}

D1={G0={x2,x3, x4}}, G1={x1}}

|L|=4

$\Delta_F(L)=16$

$\Delta F(\ell) = 3$

Функция роста

- $\Delta_F(\ell) \equiv 2^{|\ell|}$

или

- мажорируется $\Delta_F(\ell) \leq 1.5\ell^n/n!$
- n+1 минимальный объем выборки для которого нарушено условие $2^{|\ell|}$

Определения VC

- VC-измерением F(x,w) называют мощность наибольшего множества L, разбиением которого является F(x,w)

- VC-измерение комбинаторно

Скорость сходимости

- R(w) – функционал риска –вероятность ошибка классификатора.
- $R_{emp}(w)$ – функционал эмпирического риска – частота ошибка.

- При $N \to \infty$, $R_{emp}(w) \to R(w)$

Доверительный интервал

- Связь ошибки обучения, объема выборки и вероятности достижения $|R(w) - R_{emp}(w)| > \varepsilon$ для текущей модели.
- $R(w) < R_{emp}(w) + \varepsilon$

$$\varepsilon 0(N,h,a) = 2\,(h/N \times [\log(2 \times N/h)+1)] - 1/N \times \log a)$$

$$\varepsilon 1(N,h,a,R_{emp}) = 2 \times \varepsilon 0^2(N,h,a) \times (1+\sqrt{(1+R_{emp}(w)/\varepsilon 0^2(N,h,a))})$$

Если $R_{emp}(w)=0$, $\varepsilon 1(N,h,a,R_{emp}) = 4*\varepsilon 0^2(N,h,a)$

Скорость сходимости

- Ограничение 1 (общий случай)
- $R(w) \leq R_{emp}(w) + \varepsilon 1(N,h,a, R_{emp})$

- Ограничение 2 (при малой величине $R_{emp}(w)$)

$R(w) \leq R_{emp}(w) + 4\varepsilon 0(N,h,a)$

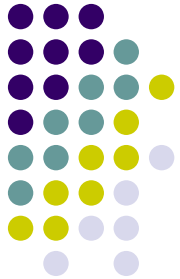
- Ограничение 3 (для большой $R_{emp}(w)$)

$R(w) \leq R_{emp}(w) + \varepsilon 0(N,h,a)$

Резюме

	малая емкость класс F(X,W)	большая емкость класс F(X,W)
Близость эмпирического правила к оптимальному	хорошая	плохая
Качество разделения	низкое	высокое

Минимизация структурного риска

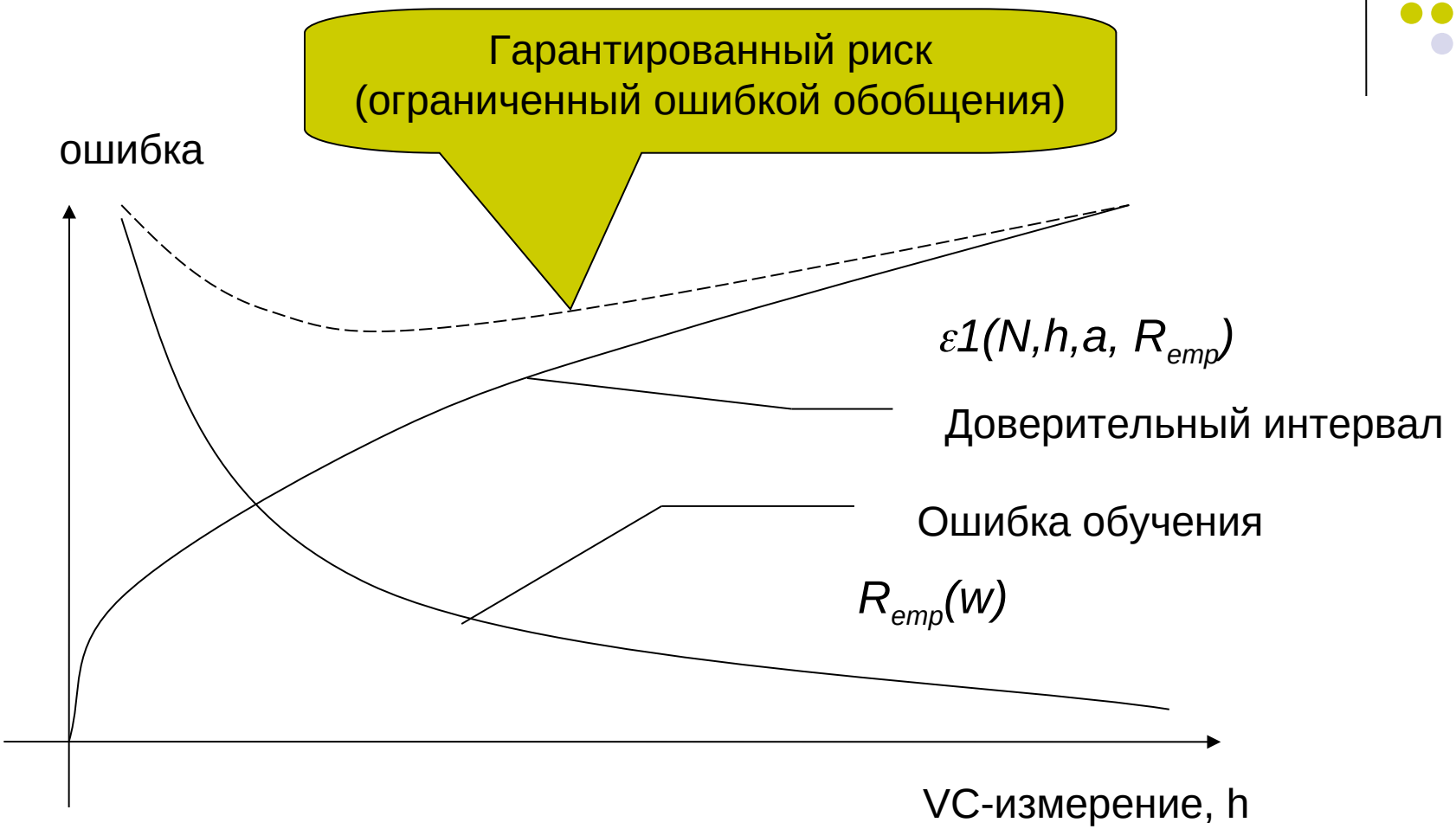


Гарантированный риск

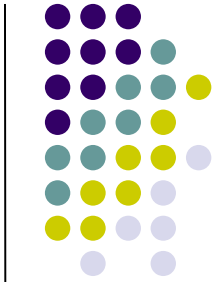
$R_{guarant}(w) = R_{emp}(w) + \varepsilon 1(N,h,a, R_{emp})$

- *h – VC* измерение классификатора *F*
- *N* – число примеров для обучения
- $\varepsilon 1$ *– доверительный интервал*

R_{gene} – ошибка обобщения


Гарантированный риск
(ограниченный ошибкой обобщения)
ошибка
ε1(N,h,a, R_emp)
Доверительный интервал
Ошибка обучения
R_emp(w)
VC-измерение, h

Требование к обучению с учителем.

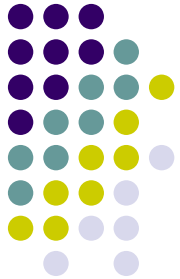


- Необходимо привести в соответствие объем классификатора и объем выборки.

- Рассмотрим
- $Fk=\{F(\boldsymbol{X},\boldsymbol{w}); w \in W\}\ k=1,2, \ldots, n$

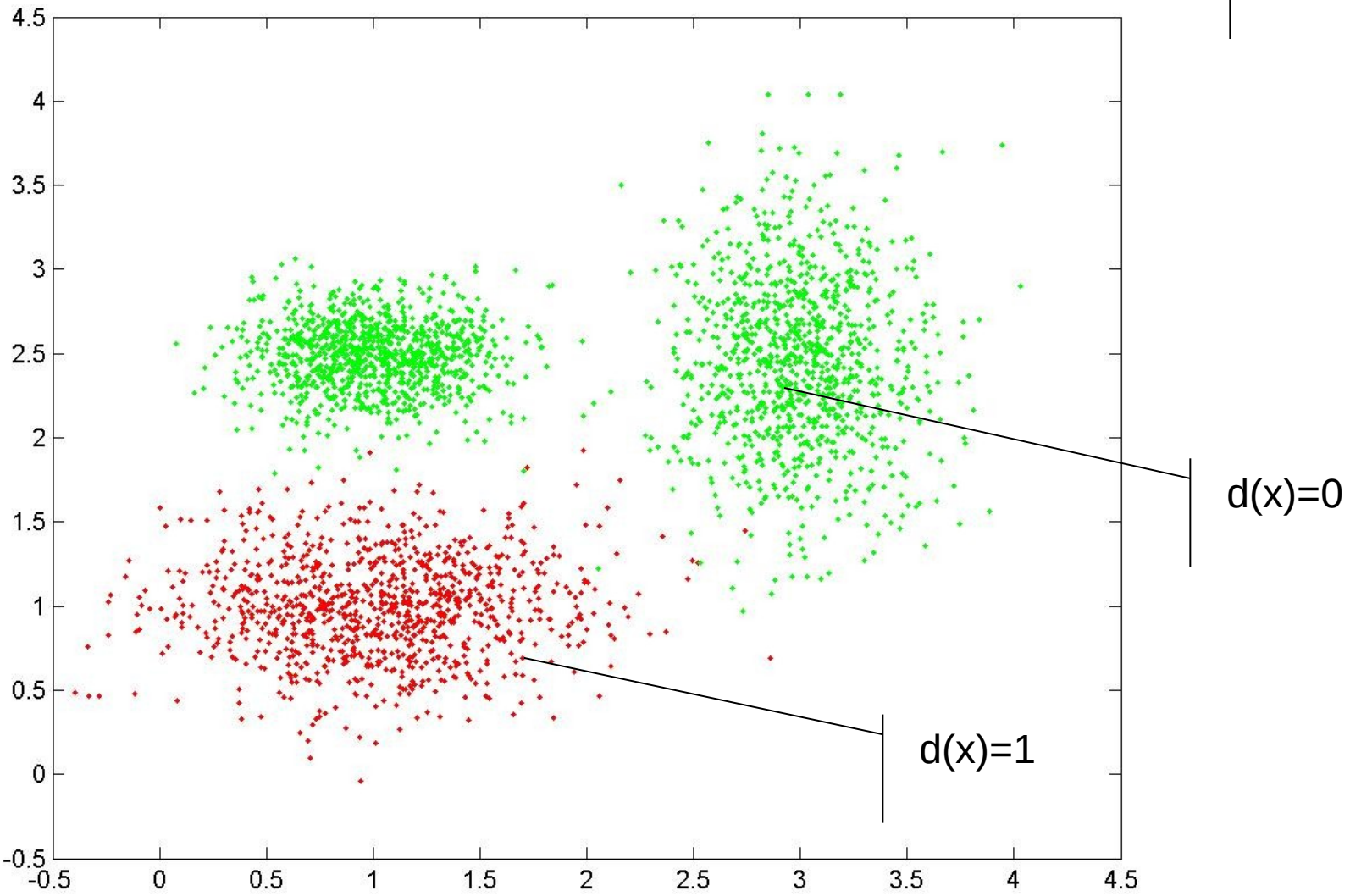
- $F1 \subset F2 \subset \ldots \subset Fn$
- $h_1 \le h_2 \le \ldots \le h_n$

Пример

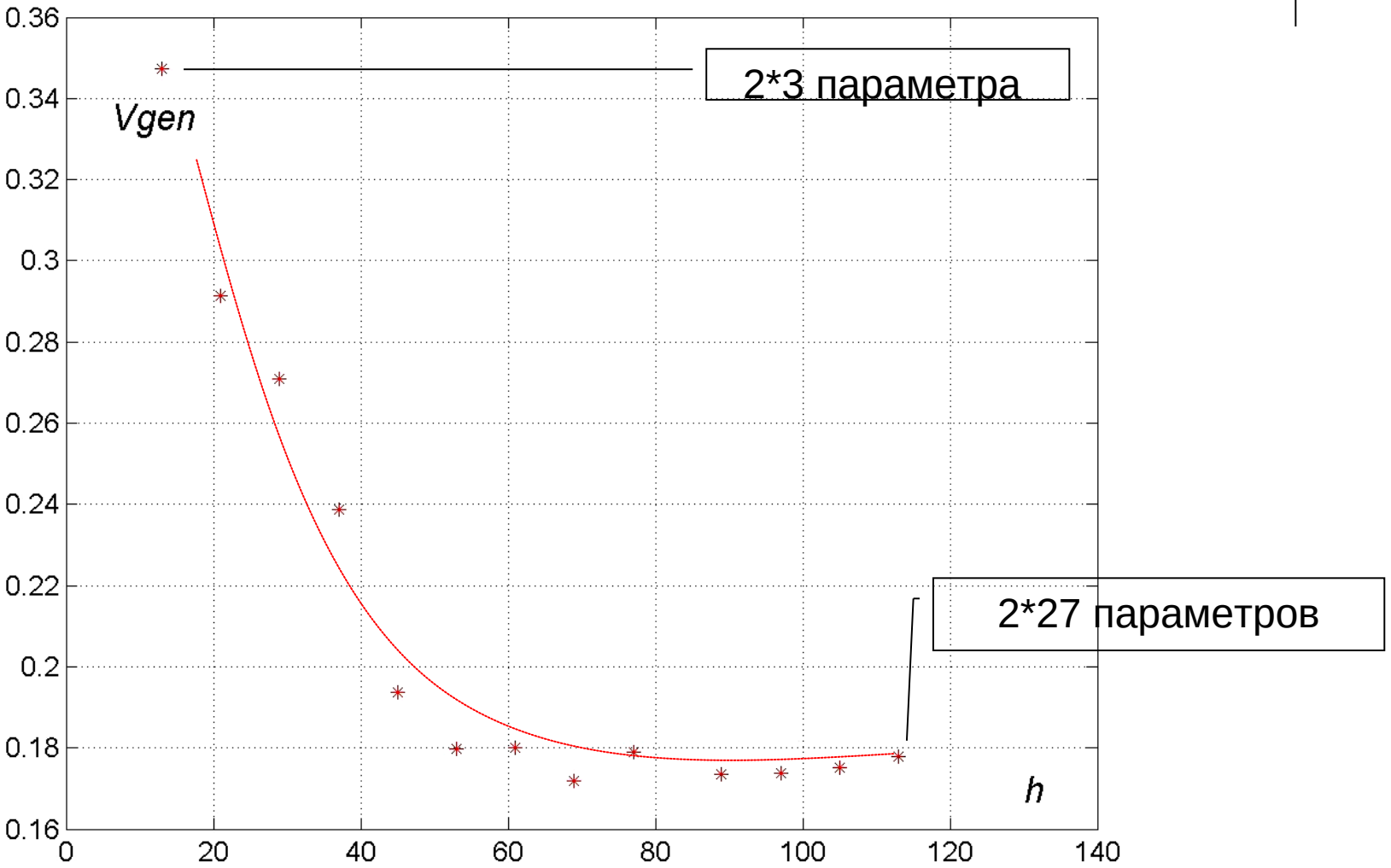


- 3000 примеров для обучения и 3000 для теста
- Определена средняя ошибка обобщения по 10 перезапускам
- Число настраиваемых параметров 2*3 (5, 7, 9, 11, 13, 15, 17, 19, 22, 24, 26, 27)

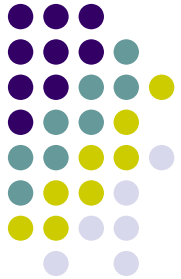
Модель множества T



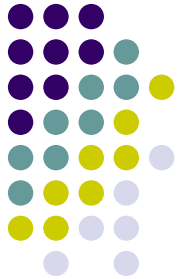
Результаты моделирования



Метод минимизации структурного риска



- Минимизируем R_{emp} *для каждого классификатора.*
- *Определяем F*, который имеет наименьший* $R_{guarant}(w)$



Резюме:

Типы обучения

- Обучение с учителем $T=\{(X_i, Y_i)\}$
- Обучение без учителя $T=\{(X_i)\}$

- ψ - стохастическое явление, где X случайный вектор из n независимых переменных
- Случайный скаляр D – зависимая переменная
- Представлено N реализаций – примеров вектора X, - {Xi}, i=1.N и соответствующих им значений случайного скаляра D, {di}

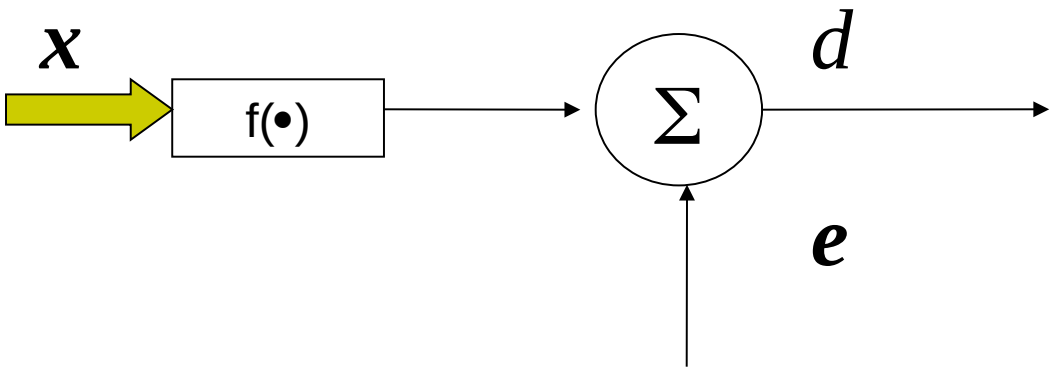
Обучающая выборка

- $T = \{(X_i, d_i)\}, \quad i = 1..N$

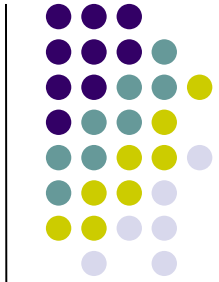
Регрессионная модель связи X и D

- $D = f(X) + \varepsilon$
- ε - ожидаемая ошибка
  - нормально распределена
  - $M(\varepsilon)=0$
- $f(\bullet)$ – детерминированная функция векторного аргумента.

Представление регрессионной модели



Свойства модели



1. Среднее значение ожидаемой ошибки ε для любой реализации X

$$E[\varepsilon|x] = 0$$

следовательно

$$f(x) = E[D|x]$$

Свойства модели

1. **Принцип ортогональности.** Ошибка ε не коррелирует с функцией регрессии $f(X)$

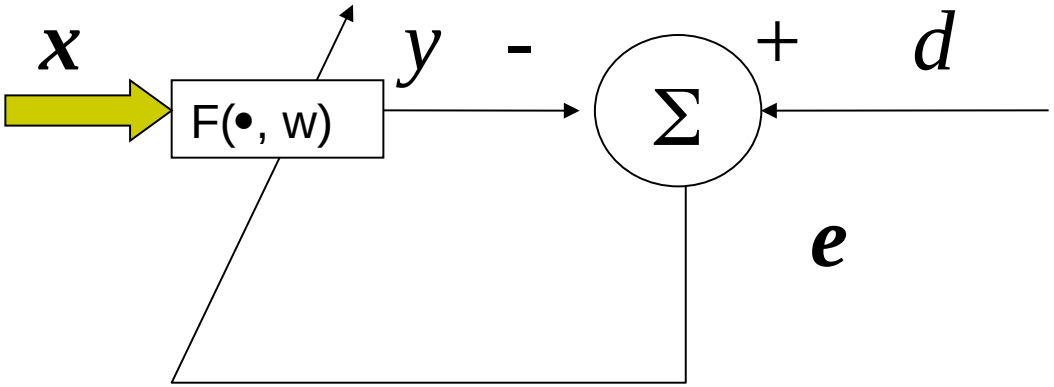
$$E[\varepsilon\, f(X)]=0$$

Вся информация о D, доступная через X, закодирована в функции регрессии $f(X)$.

Доказательство:

$$E[\varepsilon f(X)] = E[E[\varepsilon f(X)|x]] = E[f(X) E[\varepsilon|x]] = E[f(X)*0] = 0$$

Физическое представление



- Позволяет закодировать эмпирические знания выборки ***T***, с помощью синаптических весов ***w***

$$T \rightarrow w$$

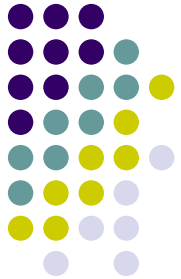
$$Y = F(X, w)$$

Определение w

- Минимизация функции стоимости

пакетное обучение:

$$Er(w) = \frac{1}{2} \Sigma (d_i - F(x_i, w))^2,$$



- Т.к. $\boldsymbol{T \rightarrow w,}$

то $F(x,w)$ заменяема на $F(x,T)$,

$Er(w) = ½\ E_T[(d_i - F(x_i,T))^2]$, где

$E_T[\bullet]$ – оператор усреднения по Т

- $d - F(x,T) = (d - f(x)) + (f(x) - F(x,T)) = \varepsilon + (f(x) - F(x,T))$,

- *тогда*

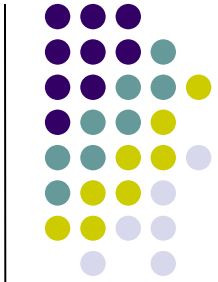
$Er(w) = \frac{1}{2} E_T[\varepsilon^2] + \frac{1}{2} E_T[(f(x) - F(x,T))^2] + E_T[\varepsilon(f(x) - F(x,T))]$, *где*

$E_T[\varepsilon (f(x) - F(x,T))]=0$

- $Er(w) = ½\ E_T[\varepsilon^2] + ½\ E_T[(f(x) - F(x,T))^2]$

- $E_T[\varepsilon^2]$ – *дисперсия ожидаемой ошибки (регрессионной) вычисляемой на T.*
- ε *- исходная* $\Rightarrow$ *не зависит от* **w,** *а значит ее можно не учитывать, т.к. Er(w)*
- $E_T[(f(x) - F(x,T))^2]$ – *среднее по ансамблю расстояние от f(x) к F(x,T)*

Функция стоимости для построения w



- $Er(w) = ½ E_T[(f(x) - F(x,T))^2]$

Мера прогнозирования

$$L_{av}(f(x), F(x,T)) = E_T[(f(x) - F(x,T))^2]$$

$$f(x) = E[D|x]$$

$$L_{av}(f(x), F(x,T)) = E_T[(E[D|X=x] - F(x,T))^2] \quad (1)$$

Ошибка оценивания регрессионной функции f(X) аппроксимационной F(x,T)

$$(E[D|X=x] - F(x,T)) =$$

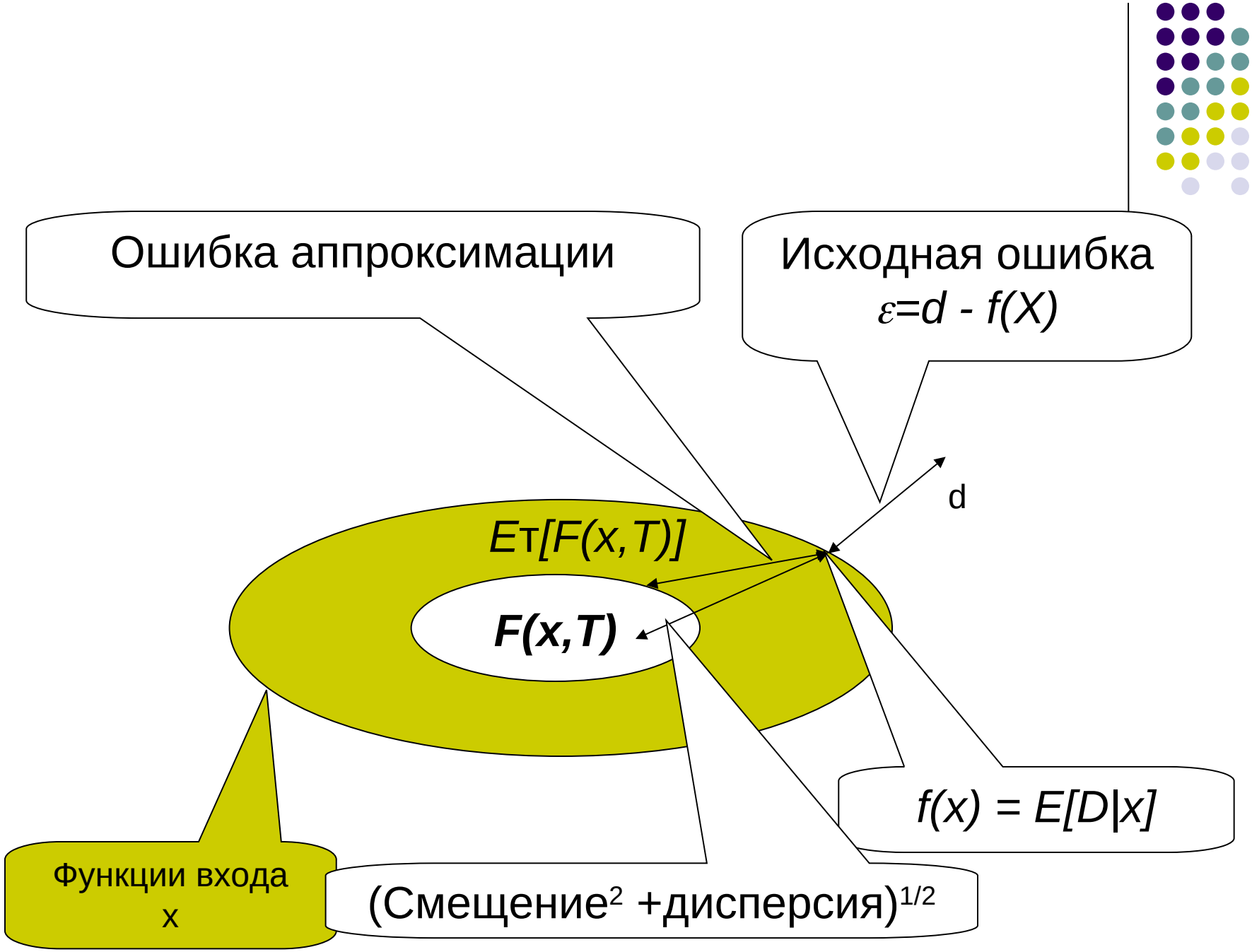
$$(E[D|X=x] - E_T[F(x,T)]) + (E_T[F(x,T)] - F(x,T)) \quad (2)$$

Тогда

$$L_{av}(f(x),F(x,T)) = E_T[(E[D|X=x] - F(x,T))^2] = B^2(w) + V(w) \quad (3)$$

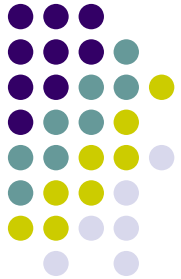
$B(w) = E_T[(E[D|X=x] - F(x,T))]$ – смещение среднего для $F(x,T)$ относительно $f(x) \Rightarrow$ ***ошибка аппроксимации***

$V(w) = E_T[(E_T[F(x,T)] - F(x,T))^2]$ – ***дисперсия*** $F(x,T)$ на всем T.


Ошибка аппроксимации
Исходная ошибка
$\varepsilon = d - f(X)$
d
$E_T[F(x,T)]$
$F(x,T)$
$f(x) = E[D|x]$
Функции входа
x
(Смещение^2 +дисперсия)$^{1/2}$

Дилемма смещения/дисперсии

- Одновременно уменьшить смещение и дисперсию можно только для бесконечно большой выборки



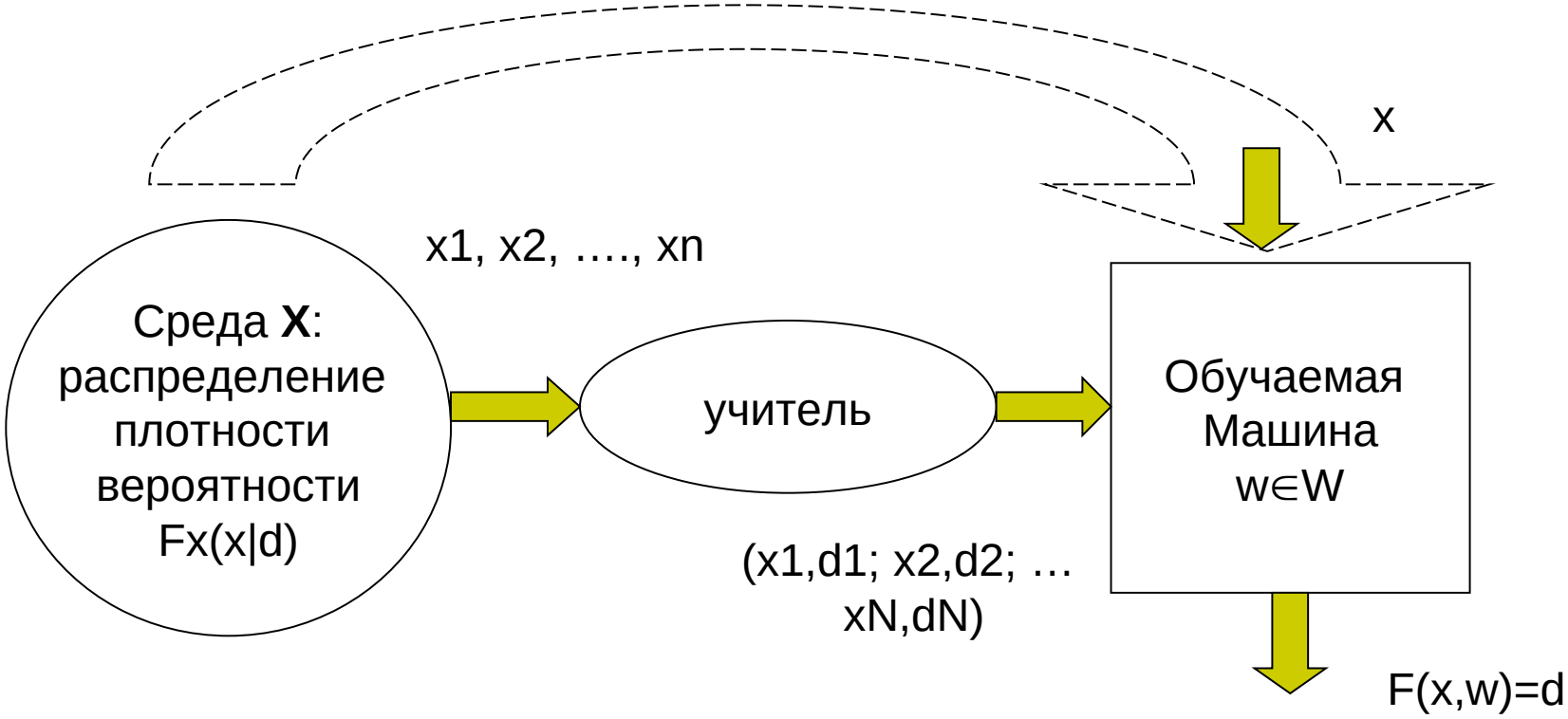
Теория статистического обучения.

Объем выборки

- **Объем выборки должен быть адекватным**

- **Каким?**

Модель обучения с учителем



Обучение с учителем

- Выбор конкретной F(x,w), который оптимально аппроксимирует отклик d, где выбор основан на N независимых, равномерно распределенных примерах.

- T = {(Xi,di)}, i=1.N

Мера потерь

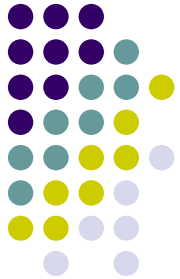
пусть

L(d, F(x,w)) – мера потерь или несходства между желаемым откликом d на вектор x и F(x,w)

$L(d, F(x,w)) = (d_i - F(x_i, T))^2$

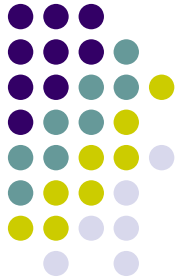
$Lav(f(x), F(x,T)) = E_T[(f(x) - F(x,T))^2]$ – *мера прогнозирования*

Ожидаемая величина потерь



- $R(w) = \int L(d, F(x,w))\, d\, F_{x,D}(x,d)$ (6)
- Цель обучения с учителем минимизация функционала риска R(w) в классе функций аппроксимации F(x,w).
- $F_{x,D}(x,d)$ обобщенная функция распределения

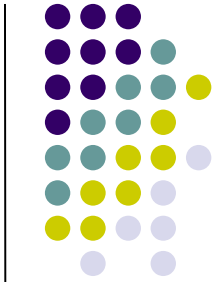
Принцип минимизации эмпирического риска



- $R_{emp}(w) = 1/N\Sigma\ L(d_i, F(x_i,w))$, i=1.N (7)

Достоинства:

- Не зависит от распределения $F_{x,D}(x,d)$
- Можно минимизировать по w



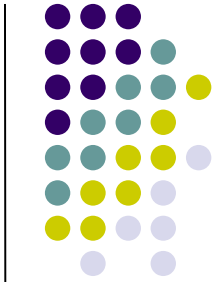
- w_{emp} и $F(x, w_{emp})$ – соответствуют минимуму функционала $R_{emp}(w)$
- w_0 и $F(x, w_0)$ - соответствуют минимуму функционала $R(w)$
- w_{emp} и $w_0 \in W$
- Найти условия, при которых $F(x, w_{emp})$ близко к $F(x, w_0)$
- Мера близости разница между $R_{emp}(w)$ и $R(w)$

- Если $R_{emp}(w)$ аппроксимирует $R(w)$ равномерно с точностью ε, то их минимумы отстоят друг от друга не более чем на 2ε

- $P(\sup | R(w) - R_{emp}(w) | > \varepsilon) \to 0 \; N \to \infty$

- $P(\sup | R(w) - R_{emp}(w) | > \varepsilon) < a$ (8)

- $P((R(w_{emp}) - R(w_0)) > 2\varepsilon) < a$ (9)



- $R(w_{emp}) - R_{emp}(w_{emp}) < \varepsilon \qquad (10)$
- $R_{emp}\,(w_0) - R\,(w_0)) < \varepsilon$
- $R_{emp}(w_{emp}) - R\,(w_0)) < 2\varepsilon \quad (11)$

Принцип минимизации эмпирического риска

1

R(w) на $R_{emp}(w)$

$R_{emp}(w) = 1/N\Sigma\ L(d_i, F(x_i,w))$

$i=1.N\ ,\ x_i,\in T$

2

Пусть w_{emp} минимизирует $R_{emp}(w)$ в W,

тогда $R_{emp}(w)$ равномерно сходится по вероятности к R(w) на растущем N.

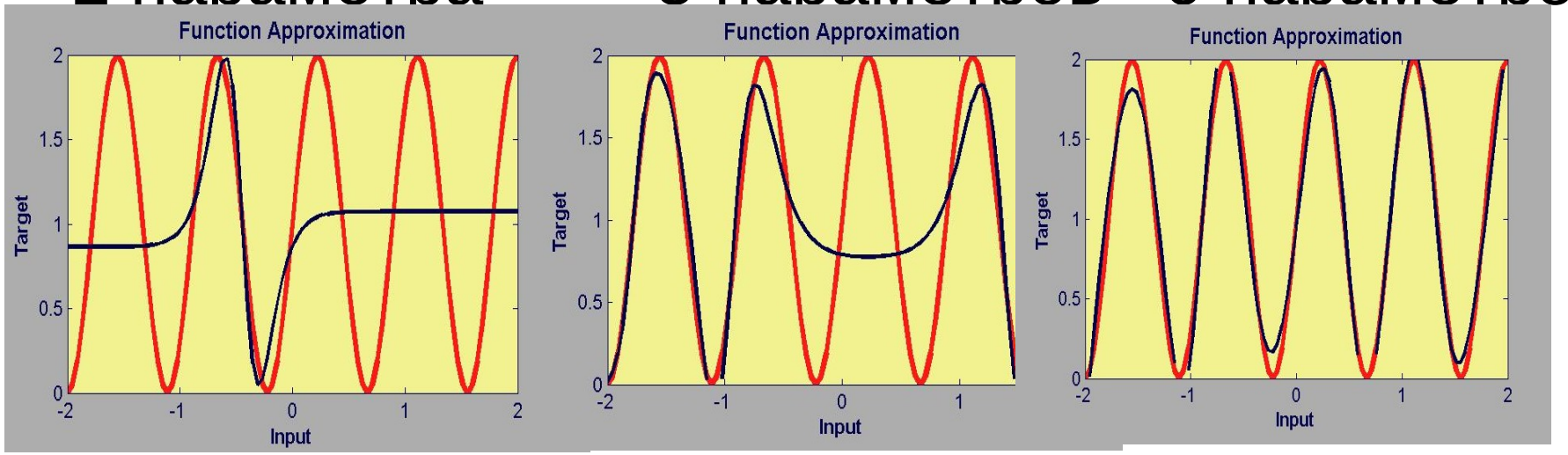
3 Равномерная сходимость

$P(\sup | R(w) - R_{emp}(w) | > \varepsilon) \rightarrow 0,\ N \rightarrow \infty$ (12)

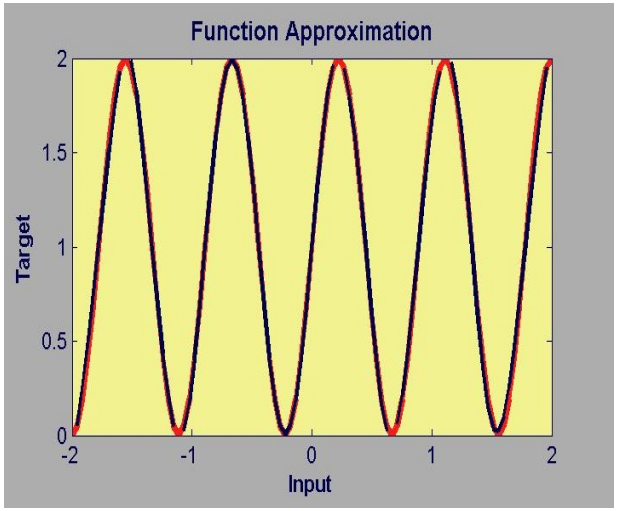
необходимое и достаточное условие непротиворечивости принципа минимизации эмпирического риска.

Выбор классификатора

- 2 параметра 5 параметров 6 параметров

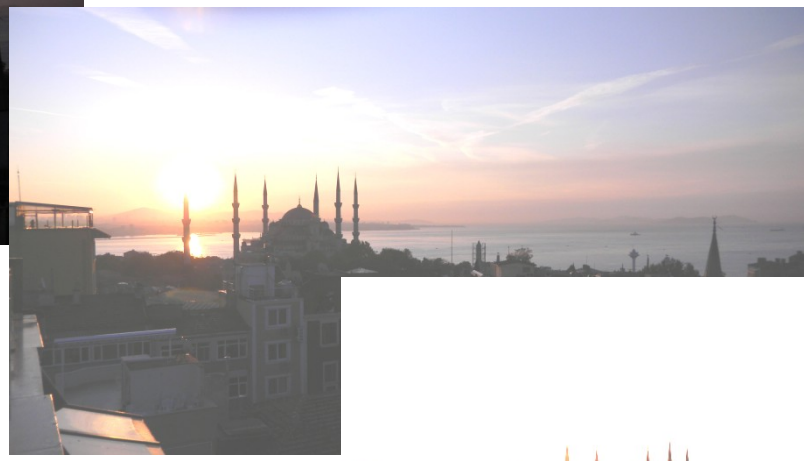


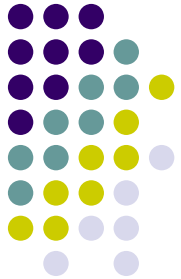
- 8 параметров



Мечеть Сулеймана

Где остановиться





Для модели с конечным VC.

- Размер выборки

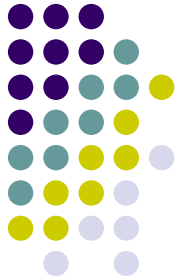
- $N=K/\varepsilon(h \log(1/\varepsilon) + \log(1/\beta))$

P(L,g|Vtrain<ε)> (1-δ) вероятность ошибки ε для алгоритма L на модели g с выборкой Vtrain

Основные положения

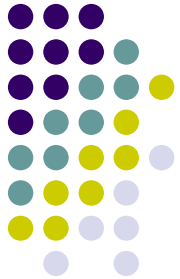
1. Множество всех объектов является вероятностным пространством с некоторой неизвестной вероятностной мерой.
2. Обучающие объекты выбираются случайно и независимо согласно этой мере.
3. Фиксируется некоторое *семейство алгоритмов*.
4. Процесс обучения заключается в построении алгоритма, принадлежащего данному семейству, и доставляющего минимум эмпирическому риску на заданной обучающей выборке.
5. Обобщающая способность алгоритма характеризуется вероятностью ошибочной классификации.
6. В общем случае мы не знаем, какой именно алгоритм будет построен в результате обучения.
7. Водится требование *равномерной сходимости* частоты ошибок к вероятности: частота ошибок должна не сильно отклоняться от их вероятности одновременно для всех алгоритмов семейства.
8. Стремление этого отклонения к нулю с ростом длины выборки принимается за определение *обучаемости семейства алгоритмов* .

Основные результаты



- Введены понятия функции роста и ёмкости семейства алгоритмов, характеризующие сложность.
- Количественные оценки, связывающие обобщающую способность алгоритмов с длиной обучающей выборки и сложностью семейства алгоритмов. Эти оценки дают достаточные условия обучаемости.
- Получены необходимые и достаточные условия *равномерной сходимости* частоты к вероятности в терминах энтропии семейства алгоритмов.
- Предложен метод структурной минимизации риска.

Основные ограничения



- **Проблема** статистической теории является завышенность оценок. Непосредственный расчёт показывает, что для надёжного обучения необходимо иметь порядка 10^6–10^8 объектов.
- **Причина** завышенности статистических оценок является их чрезмерная общность.

Модификации

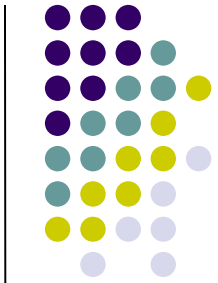
- Эффективная емкость
  - Не превосходит VC – измерение
  - учитывает особенности распределения объектов
  - Не учитывает алгоритм и выборку
  - Метод самоограничивающихся алгоритмов
- Отступы (поля - margins) – SVM
  - Альтернативная функция роста – fat-размерность
  - Обучение с максимизацией отступа

Модификации

- Композиции алгоритмов
  - Алгебраический подход к построению корректных алгоритмов
    - Области компетентности
    - Багинг- bagging
    - Бустинг -boosting

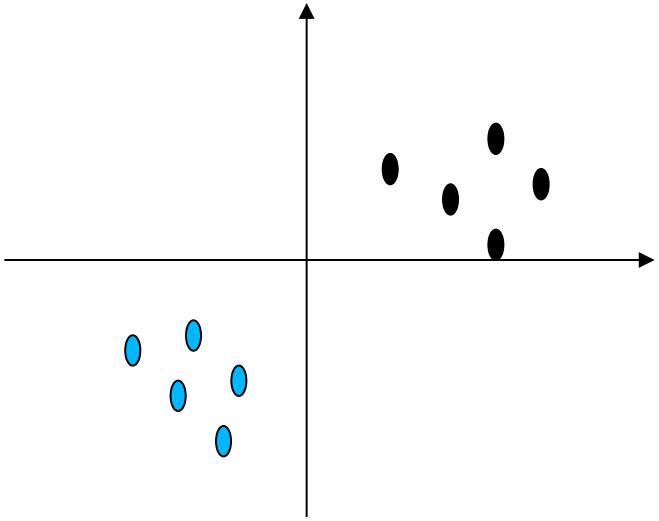
Модификации

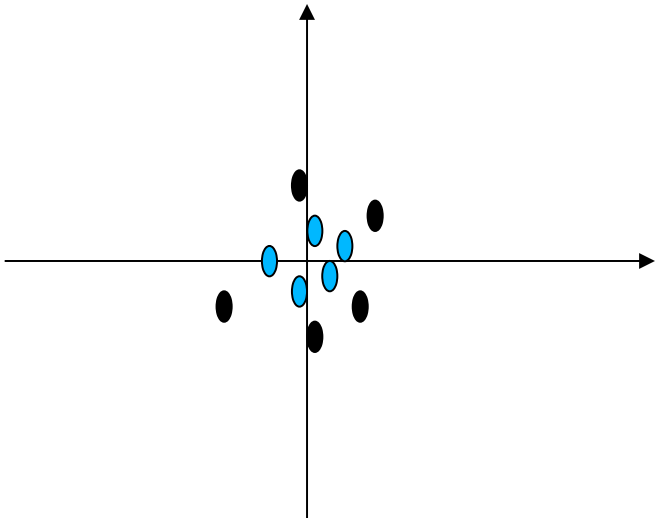
- Скользящий контроль



Определить тип классификатора

- Вариант 1
  - Линейный
  - Полином 2 степени
  - Комбинация гиперсфер
- Вариант 2
  - Линейный
  - Полином 2 степени
  - Комбинация гиперсфер





Литература

- Теория надёжности обучения по прецедентам (курс лекций, К.В.Воронцов) http://www.machinelearning.ru/wiki/index.php
- Хайкин С. Нейрокомпьютеры: полный курс. – М.:Вильямс – 2006
- Математические методы распознавания образов. Курс лекций. МГУ, ВМиК, кафедра «Математические методы прогнозирования» Местецкий Л.М., 2002–2004