Компьютерное зрение '2014
Основы машинного обучения

Who? Александр Вахитов

When? April 25, 2014

А.Н.Колмогоров

Я принадлежу к тем крайне отчаянным кибернетикам, которые не видят никаких принципиальных ограничений в кибернетическом подходе к проблеме жизни и полагают, что можно анализировать жизнь во всей ееш полноте, в том числе и человеческое сознание, методами кибернетики. Продвижение в понимании механизма высшей нервной деятельности, включая и высшие проявления человеческого творчества, по-моему, ничего не убавляет в ценности и красоте творческих достижений человека.

О чем мы поговорим

Основные понятия

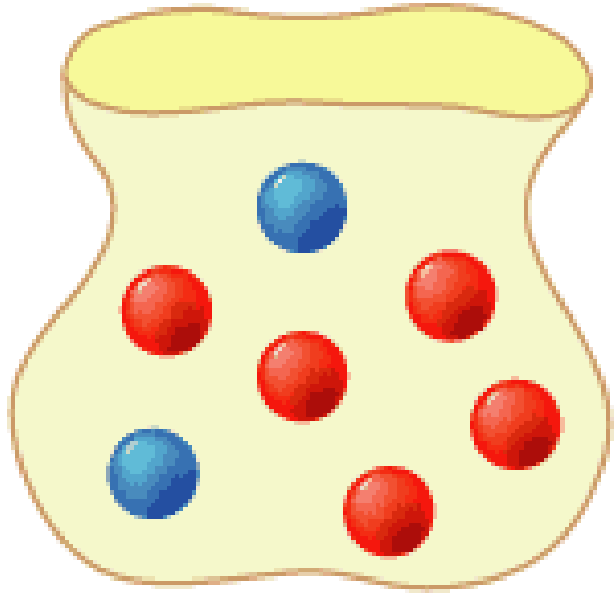
Линейные модели

Невязка и шум

Шум в наблюдениях

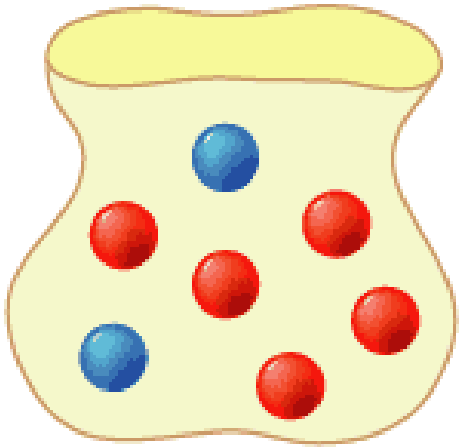
Эксперимент

Мешок с шарами 2 цветов



Достали N шаров, среди них ν - доля красных.
Задача: оценить, сколько красных в мешке?

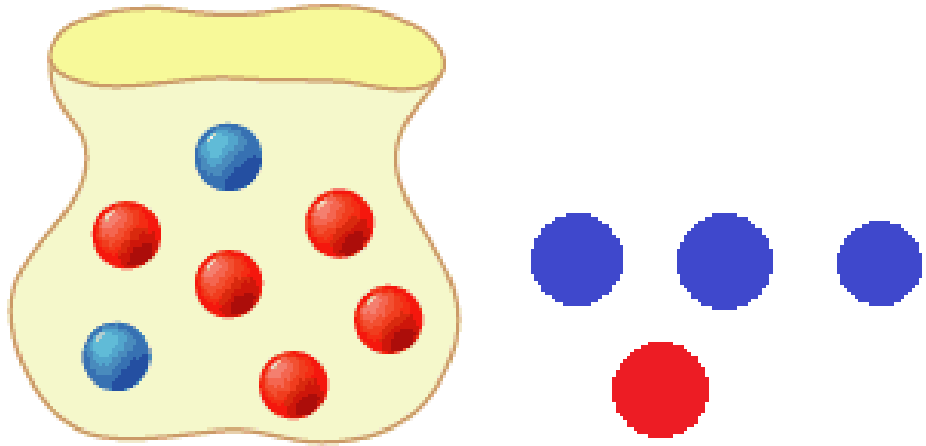
Эксперимент



$$P(red) = \mu;\ P(blue) = 1 - \mu$$

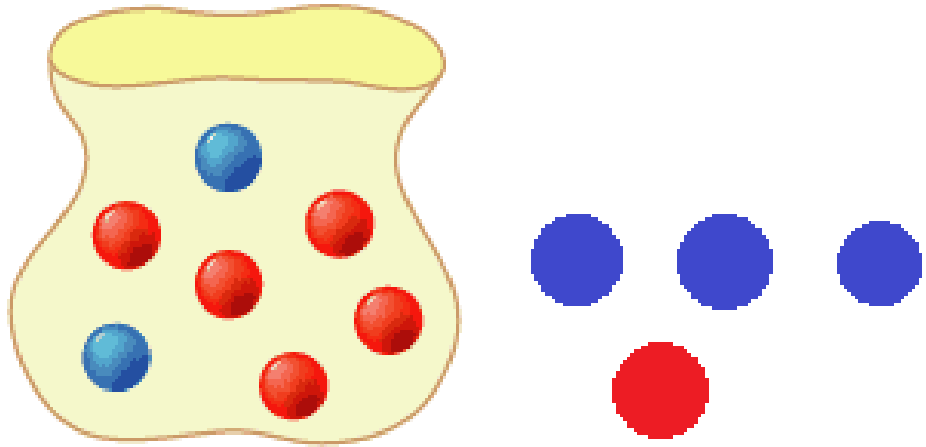
Необходимо найти μ

Эксперимент



Достали 4 шара из мешка. Что мы можем узнать о μ?

Эксперимент



Общая интуиция:

Может быть, ν сильно отличается от μ

Скорее всего, ν примерно равно μ

Неравенство Бернштейна - Хефдинга

Неравенство Бернштейна (Хефдинга, Hoeffding)
Произведем выборку длины N, при этом все случайные величины, участвующие в выборке, независимы и одинаково распределены. Если значения с.в. ограничены, то справедливо неравенство:

$$P(|\nu - \mu| > \epsilon) \leq 2e^{-2\epsilon^2 N}$$

ν - среднее по выборке размера N
μ - математическое ожидание
Иными словами, $\mu = \nu$ асимптотически ($\nu \to_{N\to\infty} \mu$)

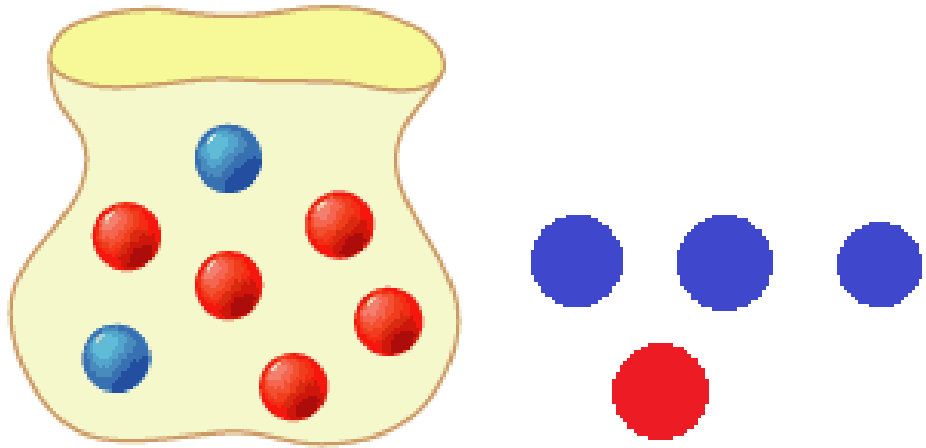
Достоинства неравенства Бернштейна - Хефдинга

$$P(|\nu - \mu| > \epsilon) \leq 2e^{-2\epsilon^2 N}$$

Правая часть не зависит от μ

Получим по выборке оценку неизвестного мат. ожидания числа шаров:

$$\nu \approx \mu \Rightarrow \mu \approx \nu$$



μ ν

Гипотеза

Шар - точка $x \in X$
$f(x)$ - неизвестная функция на множестве шаров
$h(x)$ - функция-гипотеза
Мешок для гипотезы h - цвета шаров заданы как:

- Шар синий: гипотеза h верна (для этого x, $h(x) = f(x)$)
- Шар красный: гипотеза h не верна ($h(x) \neq f(x)$)

Случайно выбирая шары из мешка, убеждаемся в истинности/ложности гипотезы для конкретного x
По выборке оцениваем долю шаров, для которых гипотеза истинна (ν)

Обучение как обобщение с выборки на множество

Задача машинного обучения - обобщение гипотезы h с выборки $x_1, x_2, \ldots, x_N$ (подмножества точек) с известными значениями f - на множество точек $x_{N+1}, \ldots$ с неизвестными значениями f.
Известно:

$$x_1, x_2, \ldots, x_N \rightarrow f(x_1), f(x_2), \ldots, f(x_N)$$

Находим:

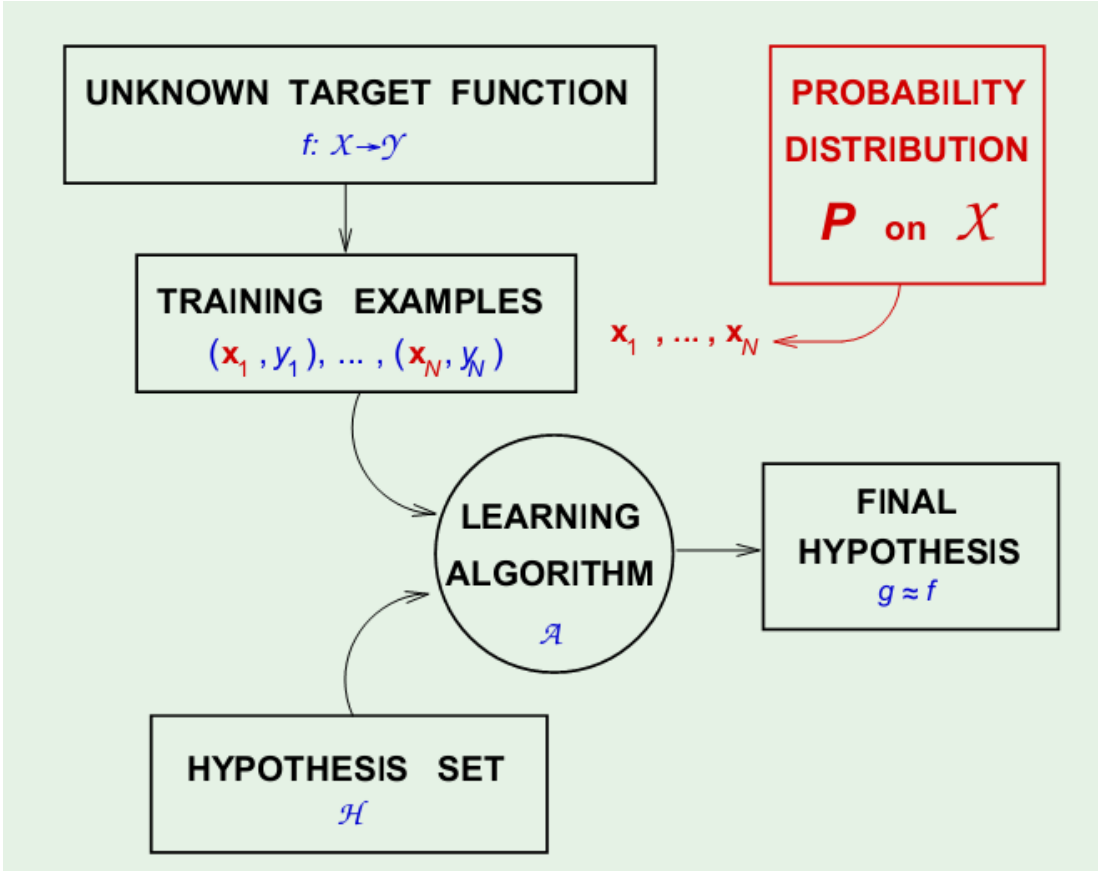
$$g: \; g(x_i) \text{ близко к } f(x_i), i = 1, \ldots, N$$

$$\Longrightarrow g(x_{N+1}) \text{ близко к } f(x_{N+1})$$

$$g(x_{N+2}) \text{ близко к } f(x_{N+2})$$

$$\ldots$$

Диаграмма обучения



(c) Y. Abu-Mostafa, Learning From Data (Caltech online)

Задача обучения

Необходимо найти наилучшее $g \in H$ из допустимого множества гипотез
Для гипотезы $h \in H$ определим

- in sample error (ошибка по выборке) $E_{in}(h) = \nu$ - доля истинных “предсказаний” значений f гипотезой h по выборке
- out of sample error (ошибка по пространству) $E_{out}(h) = \mu$ - доля истинных “предсказаний” значений f гипотезой h по всему множеству

$$P(|E_{in}(h) - E_{out}(h)| > \epsilon) \leq 2e^{-2\epsilon^2 N}$$

Немаловажная деталь

$$P(|E_{in}(h) - E_{out}(h)| > \epsilon) \leq 2e^{-2\epsilon^2 N}$$

Для каждой конкретной гипотезы неравенство выполнено.

Пусть есть множество гипотез H.

Какова вероятность, что гипотеза $g \in H$, наилучшим образом приближающая f по выборке, будет наилучшим образом приближать f по всему множеству?

Простая аналогия

С какой вероятностью монета, подброшенная 10 раз, выпадет одной и той же стороной все 10 раз?

Простая аналогия

С какой вероятностью монета, подброшенная 10 раз, выпадет одной и той же стороной все 10 раз? (0,001)

Простая аналогия

С какой вероятностью монета, подброшенная 10 раз, выпадет одной и той же стороной все 10 раз? (0,001)

Простая аналогия

С какой вероятностью монета, подброшенная 10 раз, выпадет одной и той же стороной все 10 раз? (0,001)

С какой вероятностью одна из 1000 монет, каждая из которых подброшена 10 раз, выпадет одной и той же стороной все 10 раз?

Простая аналогия

С какой вероятностью монета, подброшенная 10 раз, выпадет одной и той же стороной все 10 раз? (0,001)
С какой вероятностью одна из 1000 монет, каждая из которых подброшена 10 раз, выпадет одной и той же стороной все 10 раз?

Простая аналогия

С какой вероятностью монета, подброшенная 10 раз, выпадет одной и той же стороной все 10 раз? (0,001)
С какой вероятностью одна из 1000 монет, каждая из которых подброшена 10 раз, выпадет одной и той же стороной все 10 раз? (0,63)

Простая аналогия

С какой вероятностью монета, подброшенная 10 раз, выпадет одной и той же стороной все 10 раз? (0,001)
С какой вероятностью одна из 1000 монет, каждая из которых подброшена 10 раз, выпадет одной и той же стороной все 10 раз? (0,63)

Простое решение

$$P(|E_{in}(g) - E_{out}(g)| > \epsilon) \leq$$

$$P\big(|E_{in}(h_1) - E_{out}(h_1)| > \epsilon \text{ или}$$

$$|E_{in}(h_2) - E_{out}(h_2)| > \epsilon \text{ или } \ldots$$

$$\text{или } |E_{in}(h_M) - E_{out}(h_M)| > \epsilon\big) =$$

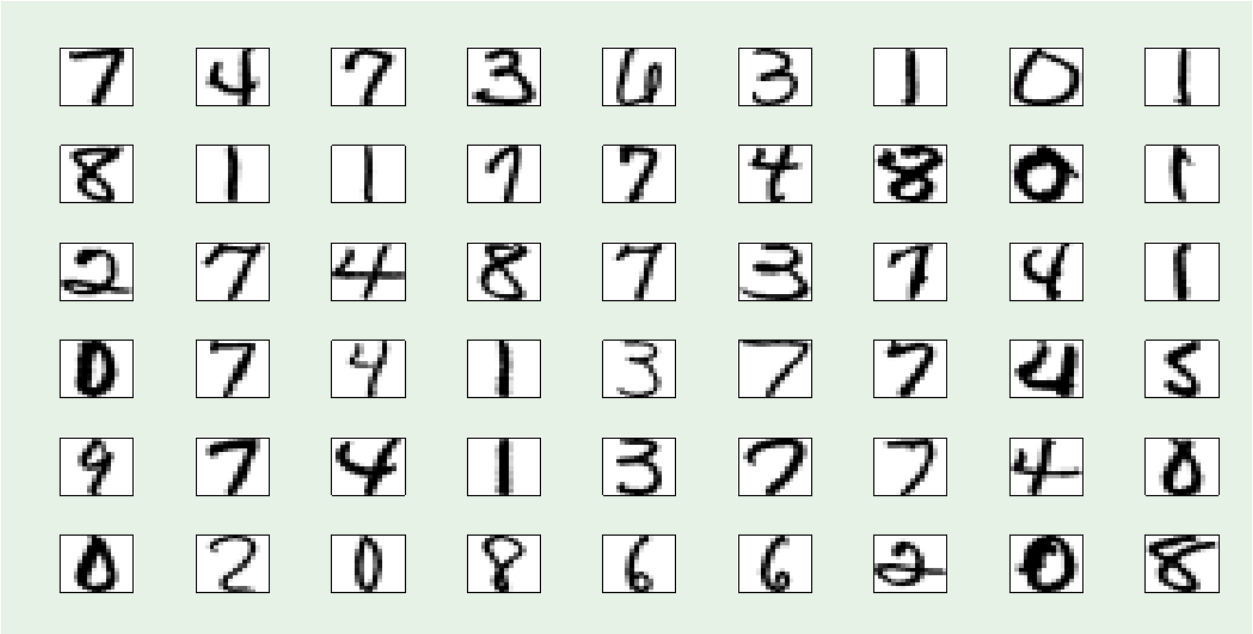
$$= \sum_{i=1}^{M} P(|E_{in}(h_i) - E_{out}(h_i)| > \epsilon)$$

$$\leq 2Me^{-2\epsilon^2 N}$$

Чем плохо?

Как улучшить?

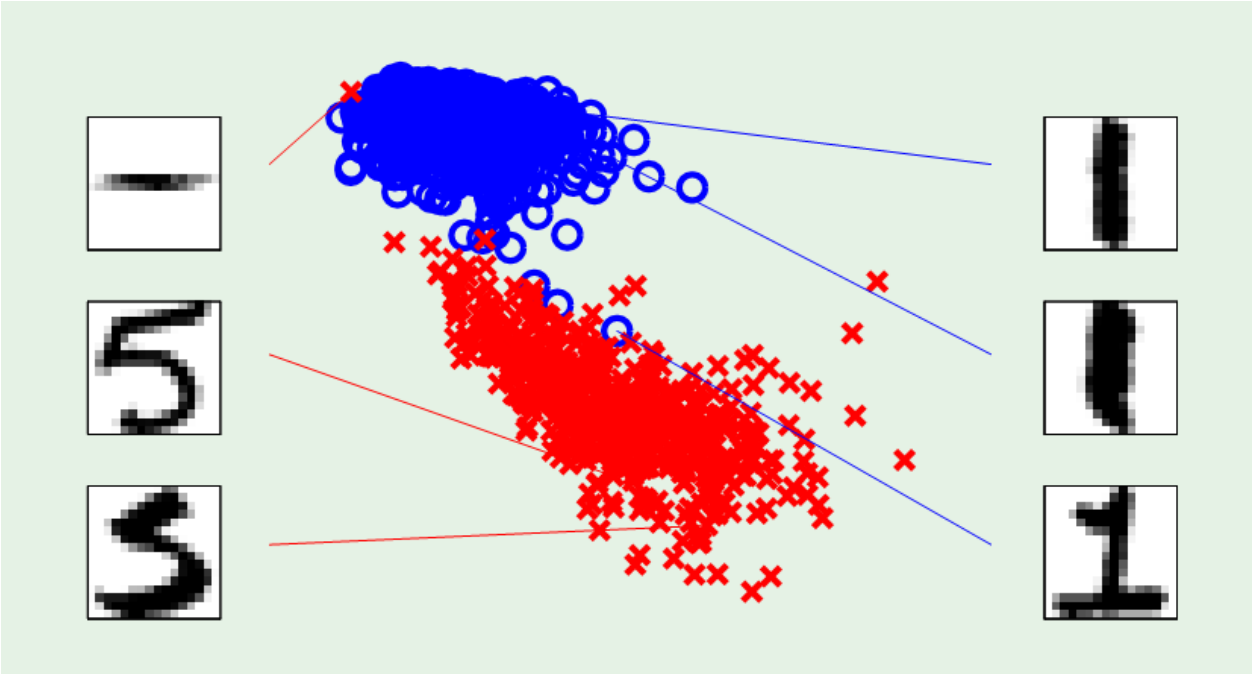
Классификация. Распознавание рукописных цифр



Выделим признаки

"Сырой" вход: значения пикселей $(x_1, \ldots, x_{256})$
"Признаки": извлечение полезной информации (интенсивность, симметрия x_1, x_2)

Пространство признаков



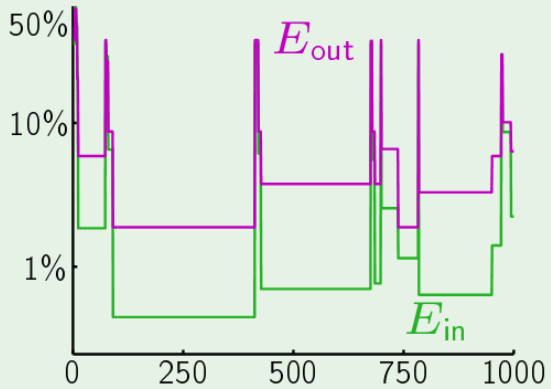
(c) Y. Abu-Mostafa, Learning From Data (Caltech online)

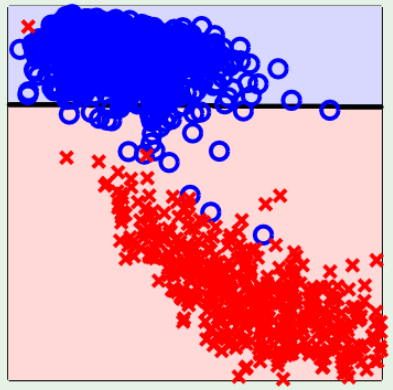
Обучение перцептрона (PLA)

разделяющая плоскость $h(x) = sign(w^T x)$

точка неверного предсказания: $sign(w^T x_n) \neq y_n$

$$w := w + y_n x_n$$





(c) Y. Abu-Mostafa, Learning From Data (Caltech online)

Линейная регрессия. Выдача кредита

Входы: возраст $x^{(1)}$, зарплата $x^{(2)}$

Выход: размер кредита

Задача: построить линейную функцию, связывающую входы и выходы

$$h(x) = \sum_{i=0}^{d} w^{(i)} x^{(i)} = w^T x$$

Задача: репликация поведения кредитного менеджера

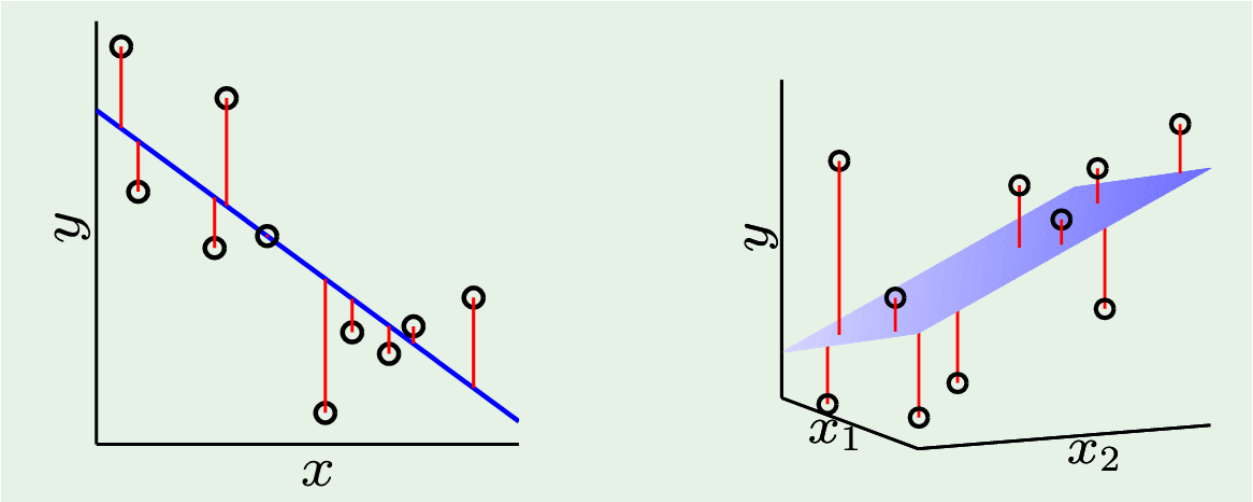
$$(x_1, y_1),\ (x_2, y_2),\ , \ldots, (x_N, y_N)$$

Измерение ошибки

Будем использовать квадратичную невязку:

$$E_{in}(h) = \frac{1}{N} \sum_{i=1}^{N} (h(x_n) - y_n)^2$$

Результат линейной регрессии



(c) Y. Abu-Mostafa, Learning From Data (Caltech online)

Результат линейной регрессии

$$\nabla E_{in}(h) = 0$$

$$E_{in}(h) = \frac{1}{N}\sum_{i=1}^{N}(h(x_n) - y_n)^2 = \frac{1}{N}\sum_{i=1}^{N}(w^T x_n - y_n)^2$$

$$= \frac{1}{N}\|Xw - y\|^2$$

Отсюда

$$\nabla E_{in}(h) = \nabla_w \frac{1}{N}\|Xw - y\|^2 = \frac{2}{N}X^T(Xw - y) = 0$$

$$X^T X w = X^T y$$

Псевдообратная матрица

A - матрица линейного преобразования из $\mathbb{R}^d$ в $\mathbb{R}^q$, $q < d$.
A^+ - псевдообратная матрица для A.
Для $v \in Ker(A), x \notin Ker(A)$ $A(x+v) = Ax$,
$A^+(Ax) = x$
Для $v \notin Ker(A)$ $Av = m$, $A^+m = v$
Через svd разложение:

$$A = U \begin{matrix} diag(\sigma_1, \ldots, \sigma_n) \\ \mathbf{0} \end{matrix} V^T,$$

$$A^+ = V \begin{matrix} diag(1/\sigma_1, \ldots, 1/\sigma_n) \\ \mathbf{0} \end{matrix} U^T,$$

Вопрос: для $v \in Ker(A)$ существует ли $w: \ A^+w = v$?

Результат линейной регрессии

$$X^T X w = X^T y$$

Если $X^T X$ обратима,

$$w = (X^T X)^{-1} X^T y$$

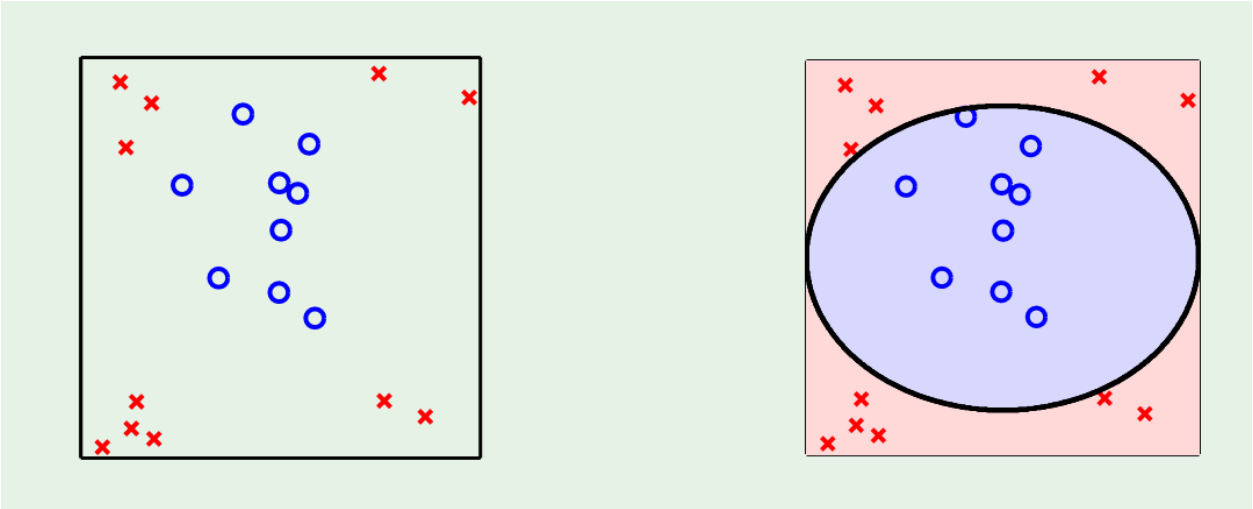
Если нет (обычно, тк измерений больше чем параметров):

$$w = (X^T X)^{+} X^T y$$

Линейная регрессия для классификации

$$y = sign(w^T x) = +1 \vee -1$$

Нелинейная классификация



(c) Y. Abu-Mostafa, Learning From Data (Caltech online)

Нелинейная классификация через линейную

$$1.\ X = (x^{(1)}, x^{(2)}) \to (x^{(1)^2}, x^{(2)^2}) = \hat{X}$$

$$2. w = (\hat{X}^T \hat{X})^+ \hat{X}^T y$$

Определение хорошей гипотезы

$f \approx h$ - что это значит?

Почти всегда невязка определяется поточечно:
$e(f(x), h(x))$

Примеры:

1. квадратичная $e(f(x), h(x)) = (f(x) - h(x))^2$
2. бинарная $e(f(x), h(x)) = id(f(x) \neq h(x))$

Полная и поточечная ошибка

Полная ошибка определяется через поточечные $e(f(x), h(x))$:

$$E_{in} = \frac{1}{N} \sum_{n=1}^{N} e(f(x_n), h(x_n))$$

$$E_{out} = \mathbb{E}_x e(f(x), h(x))$$

Классификация: как выбрать ошибку?

-	f	f	f	
h	-	+1	-1	Супермаркет vs
h	+1	true accept	false accept	
h	-1	false reject	true reject	

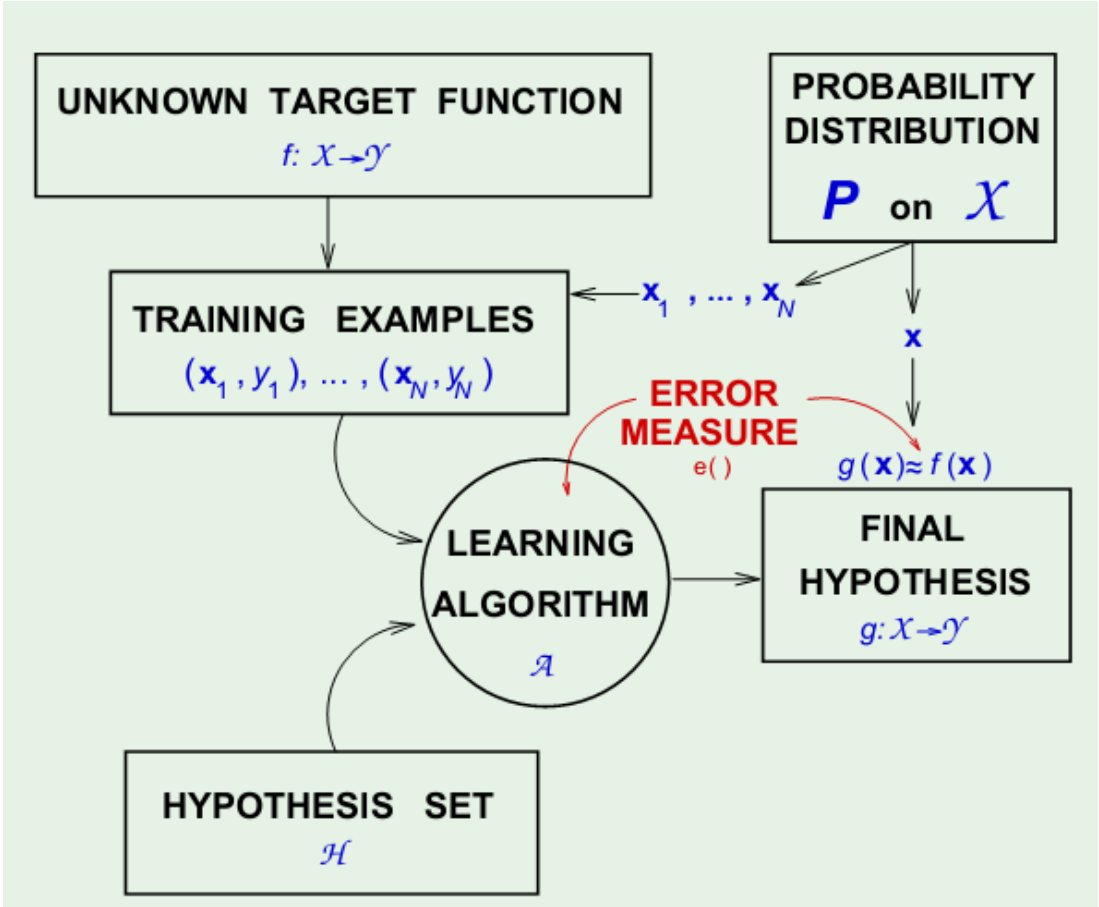
ЦРУ

Классификация: как выбрать ошибку?

-	f	f	f	
h	-	+1	-1	Супермаркет vs
h	+1	true accept	false accept	
h	-1	false reject	true reject	

ЦРУ

Роль функции ошибки



(c) Y. Abu-Mostafa, Learning From Data (Caltech online)

Задача: кредит

Входы: возраст $x^{(1)}$
зарплата $x^{(2)}$

Выход: размер кредита

Проблема входы одинаковые, выход - разный.

Что делать?

Шум в наблюдениях

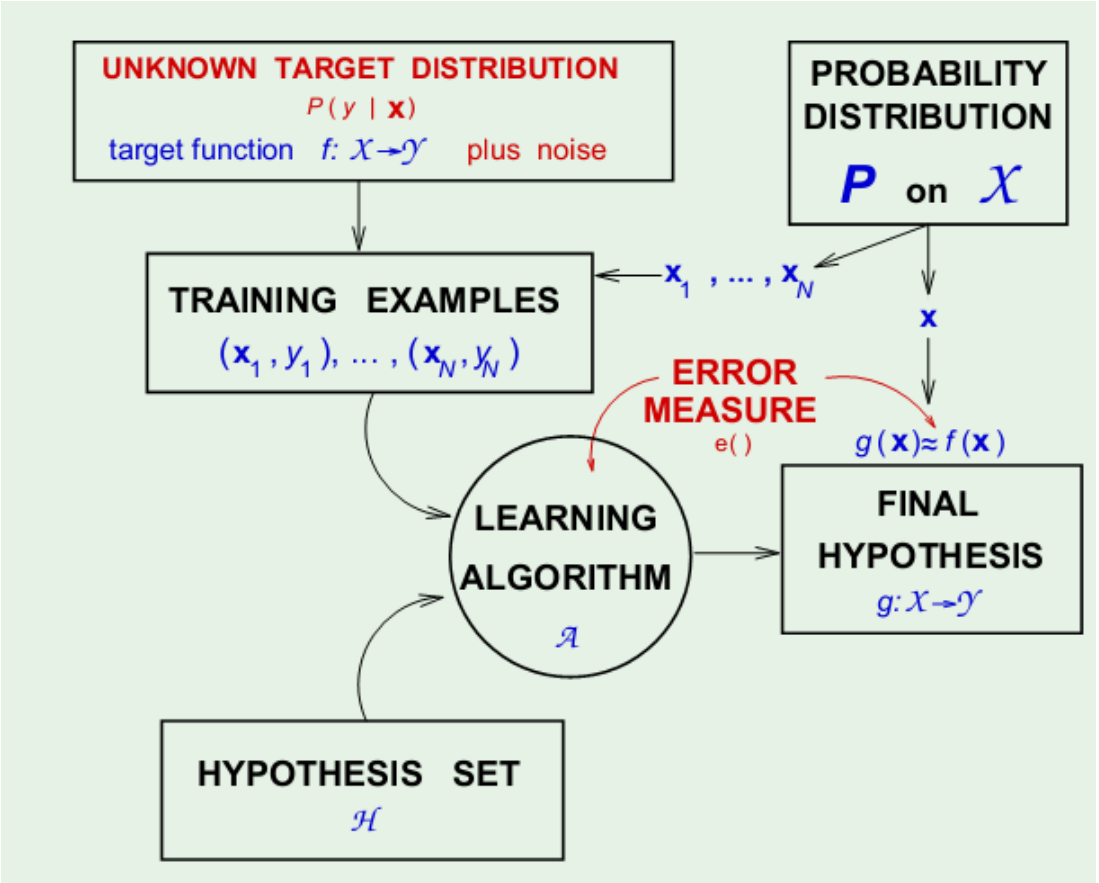
Вместо $y = f(x)$ будем использовать $P(y|x)$, так как наблюдение подвергается зашумлению
(x, y) генерируются распределением $P(y|x)P(x)$
Детерминированная компонента:

$$f(x) = \mathbb{E}(y|x)$$

Случайная компонента:

$$y - f(x)$$

Роль распределения y



(c) Y. Abu-Mostafa, Learning From Data (Caltech online)

Цель обучения в терминах вероятности

Хотим:

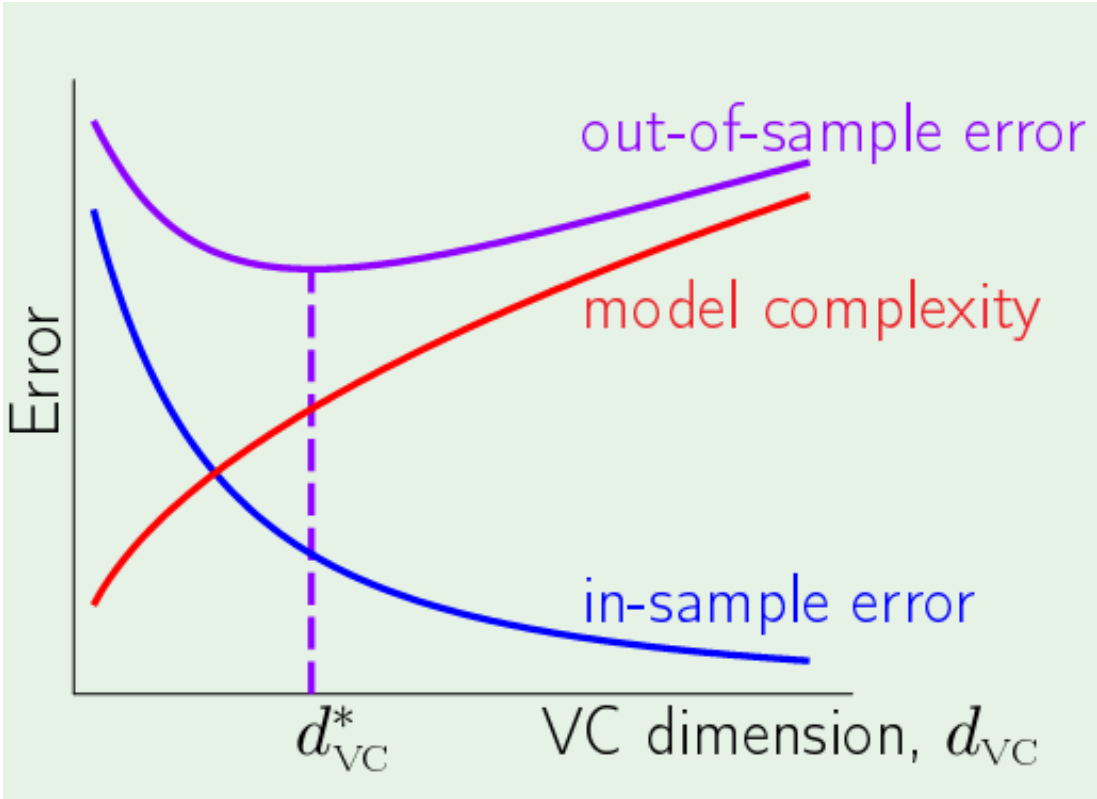
$$E_{out}(g) = 0$$

Пытаемся достичь:

$$E_{in}(g) \approx E_{out}(g) \quad E_{in}(g) = 0$$

1. Можно ли судить о E_{out} по E_{in}?
2. Как уменьшить E_{in}?

Будущая теория



Динамика ошибки от сложности модели
(c) Y. Abu-Mostafa, Learning From Data (Caltech online)