

# МАШИННОЕ ОБУЧЕНИЕ

## линейные модели классификации и регрессии

Воронцов Константин Вячеславович  
ФУПМ МФТИ • ВМК МГУ • Яндекс • FORECSYS

6 июля 2016  
Сочи, Сириус • Проектная смена • 1–24 июля 2016

- 1 Обучение как оптимизация**
  - Оптимизационные постановки задач обучения
  - Методы оптимизации
  - Линейная регрессия
- 2 Градиентные методы оптимизации**
  - Метод стохастического градиента
  - Эффект переобучение
  - Регуляризация линейных моделей
- 3 Измерение качества классификации**
  - Чувствительность и специфичность
  - ROC-кривая и площадь под кривой
  - Явная максимизация сглаженного AUC

## Восстановление зависимости по эмпирическим данным

Задача восстановления зависимости  $y = f(x)$   
по точкам *обучающей выборки*  $(x_i, y_i)$ ,  $i = 1, \dots, \ell$ .

**Дано:** векторы  $x_i = (x_i^1, \dots, x_i^n)$  — объекты обучающей выборки,  
 $y_i = f(x_i)$  — правильные ответы,  $i = 1, \dots, \ell$ :

$$\begin{pmatrix} x_1^1 & \dots & x_1^n \\ \dots & \dots & \dots \\ x_\ell^1 & \dots & x_\ell^n \end{pmatrix} \xrightarrow{f} \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix}$$

**Найти:** функцию  $a(x)$ , способную давать правильные ответы  
на *тестовых объектах*  $\tilde{x}_i = (\tilde{x}_i^1, \dots, \tilde{x}_i^n)$ ,  $i = 1, \dots, k$ :

$$\begin{pmatrix} \tilde{x}_1^1 & \dots & \tilde{x}_1^n \\ \dots & \dots & \dots \\ \tilde{x}_k^1 & \dots & \tilde{x}_k^n \end{pmatrix} \xrightarrow{a?} \begin{pmatrix} a(\tilde{x}_1) \\ \dots \\ a(\tilde{x}_k) \end{pmatrix}$$

## Типы признаков и типы задач

Типы признаков,  $x_j^i \in D_j$ , в зависимости от множества  $D_j$ :

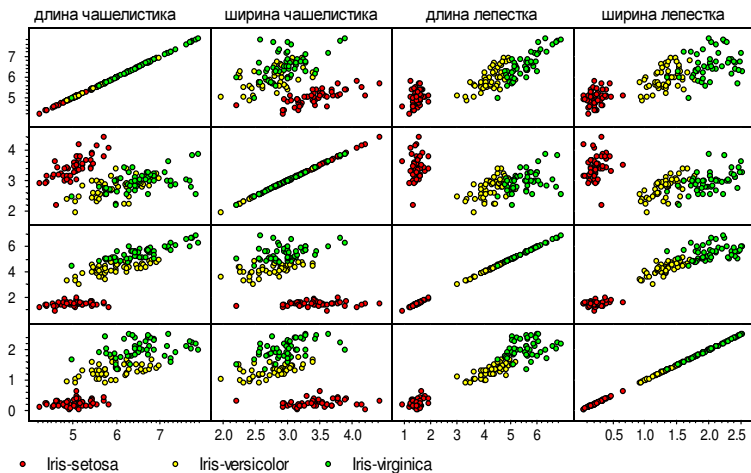
- $D_j = \{0, 1\}$  — бинарный признак;
- $|D_j| < \infty$  — номинальный признак;
- $D_j$  упорядочено — порядковый признак;
- $D_j = \mathbb{R}$  — количественный признак.

Типы задач,  $y_i \in Y$ , в зависимости от множества  $Y$ :

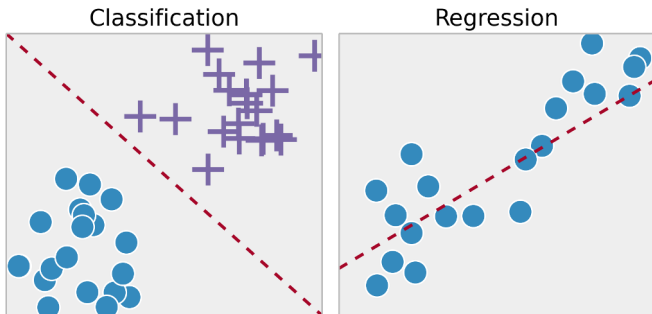
- $Y = \{0, 1\}$  или  $Y = \{-1, +1\}$  — классификация на 2 класса;
- $Y = \{1, \dots, M\}$  — на  $M$  непересекающихся классов;
- $Y = \{0, 1\}^M$  — на  $M$  классов, которые могут пересекаться;
- $Y = \mathbb{R}$  — задача восстановления регрессии;
- $Y$  упорядочено — задача ранжирования (learning to rank).

## Пример. Задача классификации цветков ириса [Фишер, 1936]

$n = 4$  признака,  $|Y| = 3$  класса, длина выборки  $\ell = 150$ .



## Линейные модели классификации и регрессии



Почему мы так любим линейные модели?

Как обучать параметры линейных моделей?

Как проверять качество предсказательной модели?

## Восстановление регрессии — это оптимизация

Задача восстановления регрессионной зависимости,  $y_i \in \mathbb{R}$

- 1 Выбираем *модель регрессии*, например, линейную:

$$a(x, w) = \langle x, w \rangle = \sum_{j=1}^n x^j w_j, \quad x, w \in \mathbb{R}^n$$

- 2 Выбираем функцию потерь, например, квадратичную:

$$\mathcal{L}(a, y) = (a - y)^2$$

- 3 Минимизируем потери *методом наименьших квадратов*:

$$Q(w) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i, w) - y_i)^2 \rightarrow \min_w$$

- 4 Проверяем прогностическую (обобщающую) способность:

$$\tilde{Q}(w) = \frac{1}{k} \sum_{i=1}^k (a(\tilde{x}_i, w) - \tilde{y}_i)^2$$

## Обучение классификации — тоже оптимизация

Задача классификации,  $y_i \in \{-1, +1\}$

- 1 Выбираем модель классификации, например, линейную:

$$a(x, w) = \text{sign}\langle x, w \rangle$$

- 2 Выбираем функцию потерь, например, бинарную:

$$\mathcal{L}(a, y) = [a(x_i, w)y_i < 0]$$

- 3 Минимизируем частоту ошибок на обучающей выборке:

$$Q(w) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i, w)y_i < 0] \rightarrow \min_w$$

- 4 Проверяем прогностическую (обобщающую) способность:

$$\tilde{Q}(w) = \frac{1}{k} \sum_{i=1}^k [a(\tilde{x}_i, w)\tilde{y}_i < 0]$$



## Обучение классификации — сглаживание функции потерь

Задача классификации,  $y_i \in \{-1, +1\}$

- 1 Выбираем модель классификации, например, линейную:

$$a(x, w) = \text{sign}\langle x, w \rangle$$

- 2 Мажорируем пороговую функцию потерь непрерывной:

$$[M_i < 0] \leq \mathcal{L}(M_i), \quad M_i = \langle x_i, w \rangle y_i \text{ — отступ (margin)}$$

- 3 Минимизируем сглаженную частоту ошибок:

$$Q(w) = \frac{1}{\ell} \sum_{i=1}^{\ell} \mathcal{L}(\langle x_i, w \rangle y_i) \rightarrow \min_w$$

- 4 Проверяем прогностическую (обобщающую) способность:

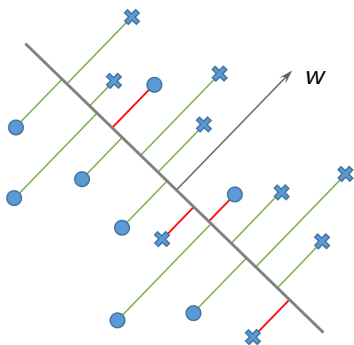
$$\tilde{Q}(w) = \frac{1}{k} \sum_{i=1}^k [\langle \tilde{x}_i, w \rangle \tilde{y}_i < 0]$$

## Геометрический смысл отступов объектов

$w$  — нормаль (направляющий вектор) разделяющей гиперплоскости, направлена в сторону класса  $+1$

$\langle x_i, w \rangle$  — проекция вектора  $x_i$  на вектор нормали  $w$

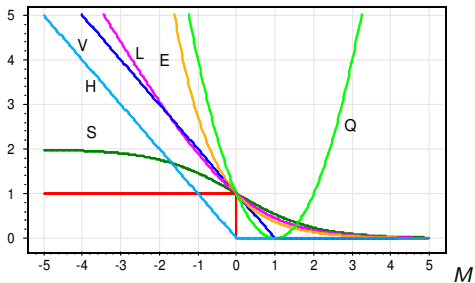
$\langle x_i, w \rangle y_i = M_i$  — проекция на нормаль в сторону класса  $y_i$



- $\times$  — объекты класса  $+1$
- $\bullet$  — объекты класса  $-1$
- $\text{---}$  — нет ошибки,  $M_i > 0$
- $\text{---}$  — ошибка,  $M_i < 0$

## Непрерывные аппроксимации пороговой функции потерь

Часто используемые непрерывные функции потерь  $\mathcal{L}(M)$ :



$$V(M) = (1 - M)_+$$

— кусочно-линейная (SVM)

$$H(M) = (-M)_+$$

— кусочно-линейная (Hebb's rule)

$$L(M) = \log_2(1 + e^{-M})$$

— логарифмическая (LR, Logistic Regression)

$$Q(M) = (1 - M)^2$$

— квадратичная (Fisher's Linear Discriminant)

$$S(M) = 2(1 + e^M)^{-1}$$

— сигмоидная (ANN, Artificial Neural Network)

$$E(M) = e^{-M}$$

— экспоненциальная (AdaBoost)

$[M < 0]$

— пороговая функция потерь.

## Общие подходы к решению оптимизационных задач

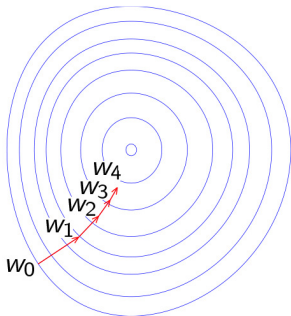
**Аналитический подход** (напр. метод наименьших квадратов):  
Если  $w$  — точка минимума *гладкой* функции  $Q(w)$ , то

$$\frac{\partial Q(w)}{\partial w_j} = 0, \quad j = 1, \dots, n.$$

Это система  $n$  уравнений с  $n$  неизвестными.

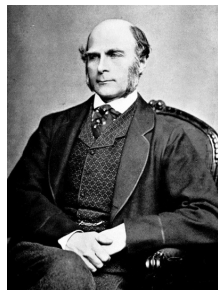
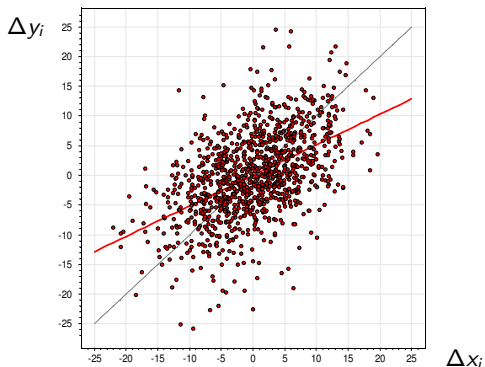
**Численный метод** — градиентный спуск:

- 1 начальное приближение  $w^0$ ,  $t := 0$ ;
- 2 **повторять**
- 3  $w_j^{t+1} := w_j^t - h^t \cdot \frac{\partial Q(w^t)}{\partial w_j}$ ,  $j = 1, \dots, n$ ;
- 4  $t := t + 1$ ;
- 5 **пока** процесс не сойдётся;



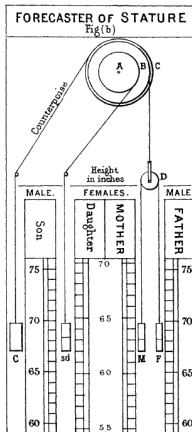
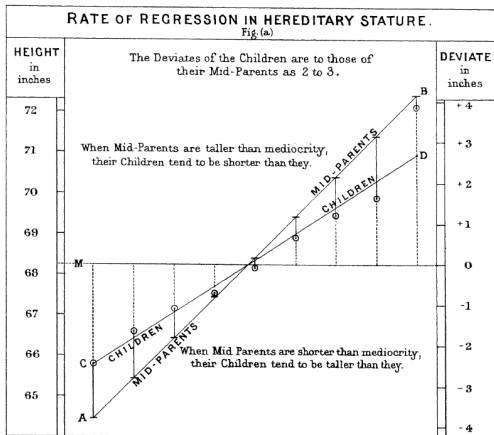
## Откуда пошло название «регрессия» (Гальтон, 1886)

Исследование наследственности роста.  
отклонение роста от среднего в популяции:  
 $\Delta x_i$  — отклонение роста отца  
 $\Delta y_i$  — отклонение роста взрослого сына



Фрэнсис Гальтон  
(1822–1911)

# Скрытый смысл: «регрессия» — сначала данные, потом модель



Galton F. Regression Towards Mediocrity in Hereditary Stature. 1886.

## Задача проведения прямой через заданные точки

**Дано:**  $x_i, y_i \in \mathbb{R}, i = 1, \dots, \ell$

**Найти:** параметры  $w = (\alpha, \beta)$  линейной модели  $y = \alpha x + \beta$

**Критерий:**  $Q(\alpha, \beta) = \sum_{i=1}^{\ell} (\alpha x_i + \beta - y_i)^2 \rightarrow \min$

Аналитический метод решения:

$$\frac{\partial Q}{\partial \alpha} = 0 \quad \Rightarrow \quad \alpha \sum_{i=1}^{\ell} x_i^2 + \beta \sum_{i=1}^{\ell} x_i - \sum_{i=1}^{\ell} x_i y_i = 0$$

$$\frac{\partial Q}{\partial \beta} = 0 \quad \Rightarrow \quad \alpha \sum_{i=1}^{\ell} x_i + \beta \sum_{i=1}^{\ell} 1 - \sum_{i=1}^{\ell} y_i = 0$$

Это система линейных уравнений  $2 \times 2$ :

$$\begin{cases} \alpha S_{xx} + \beta S_x = S_{xy} \\ \alpha S_x + \beta S_1 = S_y \end{cases} \quad \Rightarrow \quad \begin{cases} \alpha = \frac{S_{xy} S_1 - S_x S_y}{S_{xx} S_1 - S_x^2} \\ \beta = \frac{S_{xx} S_y - S_{xy} S_x}{S_{xx} S_1 - S_x^2} \end{cases}$$

## Метод стохастического градиента (SG, Stochastic Gradient)

Задача классификации:  $y_i \in \{-1, +1\}$ ,  $a(x, w) = \text{sign}\langle w, x \rangle$ .

Минимизация сглаженной частоты ошибок:

$$Q(w) = \sum_{i=1}^{\ell} \mathcal{L}(\langle w, x_i \rangle y_i) \rightarrow \min_w.$$

Один шаг *градиентного спуска*:

$$w^{t+1} := w^t - h^t \sum_{i=1}^{\ell} \mathcal{L}'(\langle w^t, x_i \rangle y_i) x_i y_i.$$

**Идея ускорения сходимости:** брать  $(x_i, y_i)$  по одному в случайном порядке и сразу обновлять вектор весов,

$$w^{t+1} := w^t - h^t \mathcal{L}'(\langle w^t, x_i \rangle y_i) x_i y_i.$$



## Алгоритм SG (Stochastic Gradient)

**Вход:** выборка  $(x_i, y_i)_{i=1}^{\ell}$ ;

**Выход:** веса  $w_1, \dots, w_n$ ;

- 1 инициализировать веса  $w_j$ ,  $j = 1, \dots, n$ ;
- 2 **повторять**
- 3 | выбрать случайный объект  $(x_i, y_i)$  из обучающей выборки;
- 4 | выбрать величину градиентного шага  $h$ ;
- 5 | выполнить градиентный шаг:  
|  $w_j := w_j - h \mathcal{L}'(\langle w, x_i \rangle y_i) x_i^j y_i$  для всех  $j = 1, \dots, n$ ;
- 6 **пока** процесс не сойдётся куда-нибудь;

**Преимущества и недостатки:**

- ⊕ можно брать какие угодно модели и функции потерь  $\mathcal{L}$
- ⊕ хорошо работает на больших выборках
- ⊖ возможно застревание в локальных экстремумах

## Эвристики

- Выбор начального приближения, например, так:

$$w_j^0 := \frac{\langle y, f_j \rangle}{\langle f_j, f_j \rangle} \quad (\text{из одномерной линейной регрессии})$$

$f_j = (x_i^j)_{i=1}^{\ell}$  — вектор значений  $j$ -го признака,  
 $y = (y_i)_{i=1}^{\ell}$  — вектор ответов.

- Выбор темпа обучения (градиентного шага)  $h^t$ :  
сходимость гарантируется для выпуклых  $Q(w)$  при

$$h^t \rightarrow 0, \quad \sum_{t=1}^{\infty} h^t = \infty, \quad \sum_{t=1}^{\infty} (h^t)^2 < \infty,$$

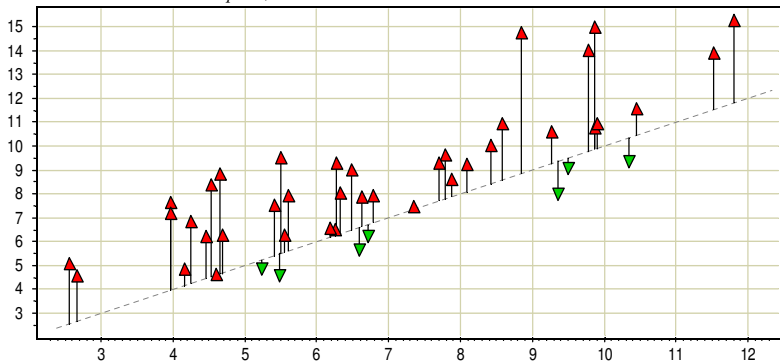
в частности можно положить  $h^t = \frac{1}{t}$ ;

- Выбор порядка предъявления объектов:
  - случайно, но попеременно из разных классов;
  - чаще брать пограничные объекты с малым  $|M_i|$ ;

## Пример. Переобучение в задаче медицинской диагностики

Задача предсказания отдалённого результата хирургического лечения атеросклероза. Точки — различные алгоритмы.

Частота ошибок на контроле, %



Частота ошибок на обучении, %

## Пример: переобучение полиномиальной регрессии

Зависимость  $y(x) = \frac{1}{1 + 25x^2}$  на отрезке  $x \in [-2, 2]$ .

Признаковое описание  $x \mapsto (1, x^1, x^2, \dots, x^n)$ .

Модель полиномиальной регрессии

$$a(x, \theta) = \theta_0 + \theta_1 x + \dots + \theta_n x^n \quad \text{— полином степени } n.$$

Обучение методом наименьших квадратов:

$$Q(a, X^\ell) = \sum_{i=1}^{\ell} (\theta_0 + \theta_1 x_i + \dots + \theta_n x_i^n - y_i)^2 \rightarrow \min_{\theta_0, \dots, \theta_n}.$$

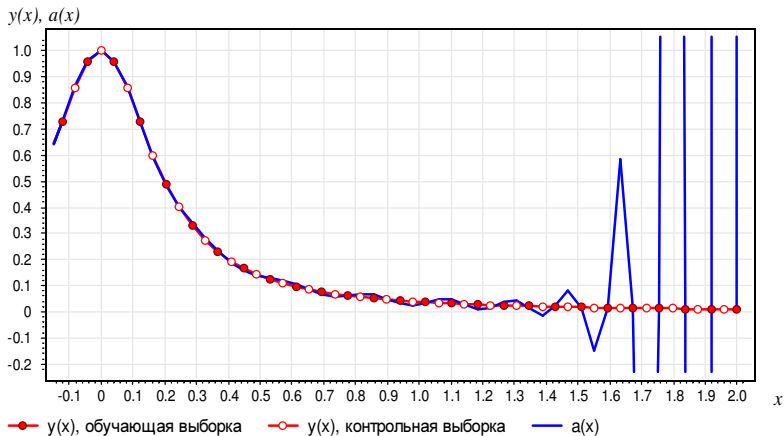
Обучающая выборка:  $X^\ell = \{x_i = 4\frac{i-1}{\ell-1} - 2 \mid i = 1, \dots, \ell\}$ .

Контрольная выборка:  $X^k = \{x_i = 4\frac{i-0.5}{\ell-1} - 2 \mid i = 1, \dots, \ell - 1\}$ .

Что происходит с  $Q(a, X^\ell)$  и  $Q(a, X^k)$  при увеличении  $n$ ?

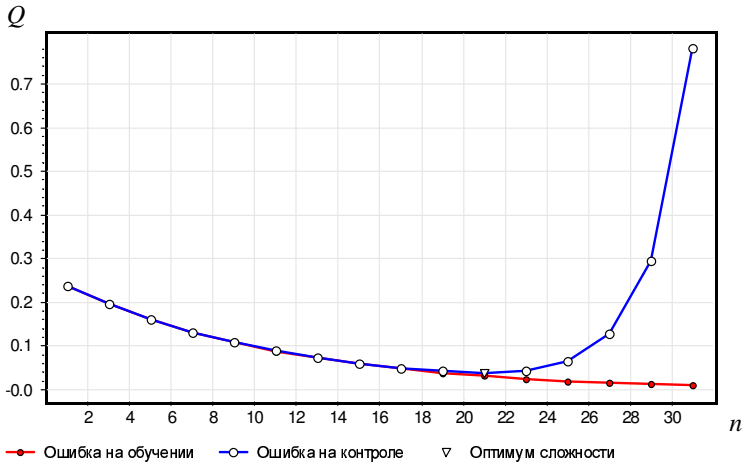
## Пример переобучения: эксперимент при $\ell = 50$ , $n = 38$

$$y(x) = \frac{1}{1 + 25x^2}; \quad a(x) \text{ — полином степени } n = 38$$



## Пример переобучения: эксперимент при $\ell = 50$ , $n = 1, \dots, 31$

Переобучение — это когда  $Q(\mu(X^\ell), X^k) \gg Q(\mu(X^\ell), X^\ell)$ :



## Эмпирические оценки обобщающей способности

Метод обучения  $\mu$  по выборке  $(x_i, y_i)_{i=1}^{\ell}$  строит алгоритм  $a$ .

- Среднее значение потерь на тестовых данных (hold-out):

$$\text{HO}(\mu, X^{\ell}, X^k) = Q(\mu(X^{\ell}), X^k) \rightarrow \min$$

- Скользящий контроль (leave-one-out),  $L = \ell + 1$ :

$$\text{LOO}(\mu, X^L) = \frac{1}{L} \sum_{i=1}^L \mathcal{L}(\mu(X^L \setminus \{x_i\}), x_i) \rightarrow \min$$

- Кросс-проверка (cross-validation) по  $N$  разбиениям,  
 $X^L = X_n^{\ell} \sqcup X_n^k$ ,  $L = \ell + k$ :

$$\text{CV}(\mu, X^L) = \frac{1}{N} \sum_{n=1}^N Q(\mu(X_n^{\ell}), X_n^k) \rightarrow \min$$

## Причины переобучения линейных моделей

- 1 слишком мало объектов; слишком много признаков;
- 2 линейная зависимость (мультиколлинеарность) признаков:  
пусть построен классификатор:  $a(x, w) = \text{sign}\langle x, w \rangle$ ;  
мультиколлинеарность:  $\exists v \in \mathbb{R}^n: \forall x \langle x, v \rangle \approx 0$ ;  
тогда  $\forall \gamma \in \mathbb{R} \quad a(x, w) \approx \text{sign}\langle x, w + \gamma v \rangle$

### Последствия:

- решение неединственно и неустойчиво;
- появляются слишком большие веса  $+w_j$  или  $-w_j$ ;
- $Q(w)$  на обучении много меньше, чем на контроле;

Спасает *регуляризация* — введение дополнительного критерия:

$$\|w\|^2 = \sum_{j=1}^n w_j^2 \rightarrow \min.$$



## Метод сокращения весов (weight decay)

Штраф за увеличение нормы вектора весов:

$$Q_{\tau}(w) = Q(w) + \frac{\tau}{2} \sum_{j=1}^n w_j^2 \rightarrow \min_w.$$

Компоненты вектора градиента:

$$\frac{\partial}{\partial w_j} Q_{\tau}(w) = \frac{\partial}{\partial w_j} Q(w) + \tau w_j, \quad j = 1, \dots, n.$$

Модификация градиентного шага:

$$w_j^{t+1} := w_j^t (1 - h^t \tau) - h^t \frac{\partial}{\partial w_j} Q(w^t), \quad j = 1, \dots, n.$$

Параметр регуляризации  $\tau$  подбирается экспериментально, по качеству на контрольной выборке.

## Терминология, пришедшая из медицинской диагностики

*Положительный диагноз* — алгоритм предсказывает болезнь (хотя, казалось бы, что тут положительного...)

Доля больных с верным положительным диагнозом:

$$\text{чувствительность} = \frac{\sum_{i=1}^{\ell} [y_i = +1][a(x_i) = +1]}{\sum_{i=1}^{\ell} [y_i = +1]}$$

Доля здоровых с верным отрицательным диагнозом:

$$\text{специфичность} = \frac{\sum_{i=1}^{\ell} [y_i = -1][a(x_i) = -1]}{\sum_{i=1}^{\ell} [y_i = -1]}$$

Чувствительность и специфичность хотим максимизировать.

- ⊕ Они не зависят от соотношения мощностей классов.
- ⊕ Хорошо подходят для несбалансированных выборок.

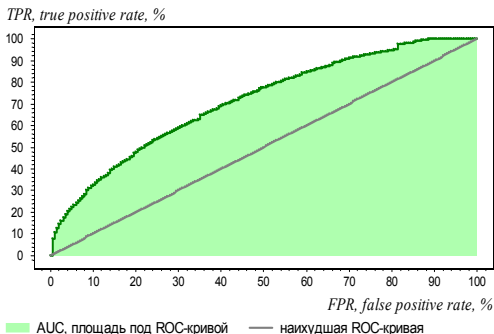
## Определение ROC-кривой

Модель классификации:  $a(x) = \text{sign}(\langle x, w \rangle - w_0)$

по оси X:  $1 - \text{специфичность} = \text{FPR}$ , false positive rate,

по оси Y: чувствительность = TPR, true positive rate

Каждая точка ROC-кривой соответствует значению порога  $w_0$   
(ROC — «receiver operating characteristic»),



## AUC (area under curve) — площадь под ROC-кривой

Модель классификации:  $a(x) = \text{sign}(\langle x, w \rangle - w_0)$

Доля правильно упорядоченных пар объектов  $(x_i, x_s)$  из разных классов,  $y_i = -1$ ,  $y_s = +1$  (докажите!):

$$\text{AUC} = \frac{\sum_{i=1}^{\ell} \sum_{s=1}^{\ell} [y_i < y_s] [\langle x_i, w \rangle < \langle x_s, w \rangle]}{\sum_{i=1}^{\ell} \sum_{s=1}^{\ell} [y_i < y_s]}$$

**Преимущества AUC:**

- ⊕ не зависит от порога  $w_0$ , оценивает только качество  $w$ ;
- ⊕ не зависит от численности классов;
- ⊕ это общепринятая мера качества классификации;

Чтобы измерить предсказательную способность метода обучения  $\mu$ , надо вычислять AUC на контрольной выборке.

## Явная максимизация сглаженного AUC

Максимизация доли правильно упорядоченных пар  $(x_i, x_s)$ :

$$\text{AUC} = \sum_{i,s: y_i < y_s} [\langle x_i, w \rangle < \langle x_s, w \rangle] \rightarrow \max_w$$

Максимизация сглаженного AUC:

$$Q(w) = \sum_{i,s: y_i < y_s} \mathcal{L}(\underbrace{\langle x_s, w \rangle - \langle x_i, w \rangle}_{M_{is}}) \rightarrow \min_w$$

$\mathcal{L}(M)$  — гладкая убывающая функция отступа,

$M_{is}$  — новое понятие отступа, теперь для пар объектов.

Метод стохастического градиента по парам  $(x_i, x_s)$ :  $y_i < y_s$

$$w^{t+1} := w^t - h^t \mathcal{L}'(\langle x_s - x_i, w^t \rangle)(x_s - x_i)$$

## Метод стохастического градиента для максимизации AUC

**Вход:** выборка  $(x_i, y_i)_{i=1}^{\ell}$ ,  $y_i \in \{-1, +1\}$ , параметры  $h, \tau$ ;

**Выход:** веса  $w_1, \dots, w_n$ ;

1 инициализировать веса  $w_j = \frac{\langle y, f_j \rangle}{\langle f_j, f_j \rangle}$ ,  $j = 1, \dots, n$ ;

2 **повторять**

3 | выбрать случайную пару обучающих объектов  $x_i, x_s$   
из разных классов:  $y_i = -1$ ,  $y_s = 1$ ;

4 | выбрать величину градиентного шага  $h$ ;

5 | выполнить градиентный шаг:

$$M_{is} := \langle x_s - x_i, w \rangle;$$

**для всех**  $j = 1, \dots, n$

$$w_j := w_j(1 - \tau h) + h \mathcal{L}'(M_{is})(x_s^j - x_i^j);$$

6 **пока** процесс не сойдётся куда-нибудь;

## Резюме

- Обучение — это оптимизации (почти во всех методах)
- Лучшие методы классификации основаны на сглаживании пороговой функции потерь
- Метод наименьших квадратов для линейной регрессии сводится к решению системы линейных уравнений  $n \times n$
- Метод стохастического градиента позволяет единообразно решать самые разные задачи, в том числе для Big Data
- Регуляризация помогает против переобучения
- Можно максимизировать непосредственно AUC

Воронцов Константин Вячеславович

[voron@forecsys.ru](mailto:voron@forecsys.ru)

[www.MachineLearning.ru](http://www.MachineLearning.ru) • Участник:Vokov

Если что-то было не понятно,  
не стесняйтесь подходить и спрашивать :)