Курс «Введение в машинное обучение»

Метод стохастического градиента и линейные модели

Воронцов Константин Вячеславович
k.v.vorontsov@phystech.edu
http://www.MachineLearning.ru/wiki?title=User:Vokov

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Введение в машинное обучение (курс лекций, К.В.Воронцов)»

МФТИ.ФПМИ.ИС.ИАД • 20 февраля 2025

Содержание

Общая постановка большинства задач машинного обучения

Дано: X — пространство *объектов*
$X^\ell = \{x_1, \ldots, x_\ell\} \subset X$ — *обучающая выборка* (training sample)
$a(x, w)$, $a\colon X \times W \to Y$ — параметрическая модель, гипотеза

Найти $w \in W \subseteq \mathbb{R}^N$ — вектор параметров модели $a(x, w)$

Критерий минимизации эмпирического риска
(empirical risk minimization, ERM):

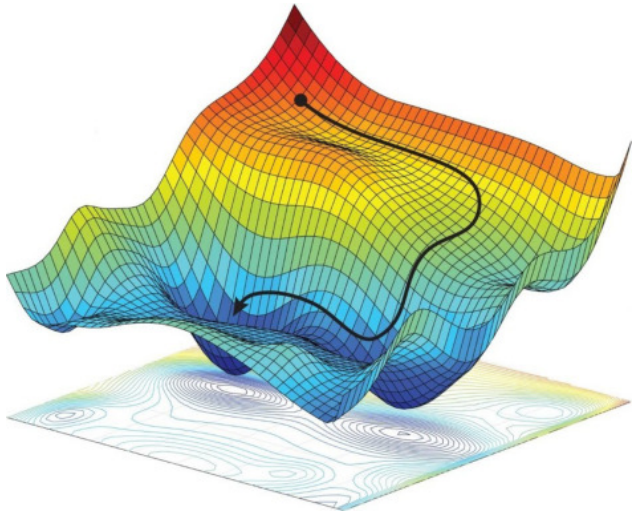
$$Q(w) = \frac{1}{\ell} \sum_{i=1}^{\ell} \mathscr{L}(w, x_i) + \tau \mathscr{R}(w) \to \min_w$$

$\mathscr{L}(w, x)$ — *функция потерь* (loss function),
тем больше, чем хуже ответ модели $a(x, w)$ на объекте x

$\mathscr{R}(w)$ — *регуляризатор* для формализации дополнительных требований к модели, τ — коэффициент регуляризации

Градиентный метод численной оптимизации

$$Q(w) = \frac{1}{\ell}\sum_{i=1}^{\ell}\mathscr{L}(w, x_i) + \tau\mathscr{R}(w) \to \min_w$$



Метод *градиентного спуска*:

$w^{(0)} :=$ начальное приближение;

$w^{(t+1)} := w^{(t)} - h\nabla Q(w^{(t)})$

где $\nabla Q(w) = \left(\frac{\partial Q(w)}{\partial w_j}\right)_{j=1}^{N}$ — *вектор градиента*,

h — *градиентный шаг*, называемый также *темпом обучения*

$$w^{(t+1)} := w^{(t)} - h\frac{1}{\ell}\sum_{i=1}^{\ell}\nabla\Big(\mathscr{L}\big(w^{(t)}, x_i\big) + \tau\mathscr{R}(w)\Big)$$

Идея ускорения сходимости:

брать объекты x_i по одному и сразу обновлять вектор весов

Метод стохастического градиента SG (Stochastic Gradient)

Задача ERM: $Q(w) = \frac{1}{\ell}\sum_{i=1}^{\ell}\mathscr{L}(w, x_i) + \tau\mathscr{R}(w) \to \min_w$

Вход: выборка X^ℓ, параметры h, τ, λ;
Выход: вектор весов w;

инициализировать веса w_j, $j = 1, \ldots, N$;
инициализировать оценку $Q(w)$ по небольшой подвыборке;
повторять
  объект x_i выбрать из X^ℓ случайным образом;
  потеря: $\varepsilon_i := \mathscr{L}(w, x_i)$;
  градиентный шаг: $w := w - h\nabla\mathscr{L}(w, x_i) - h\tau\nabla\mathscr{R}(w)$;
  рекуррентная оценка критерия: $Q := \lambda\varepsilon_i + (1-\lambda)Q$;
пока значение Q и/или веса w не сойдутся;

H. Robbins, S. Monro. A stochastic approximation method. 1951.

Откуда взялась такая рекуррентная оценка функционала?

Проблема: вычисление оценки Q по всей выборке $x_1, \ldots, x_\ell$ намного дольше градиентного шага по одному объекту x_i.

Решение: использовать приближённую рекуррентную формулу.

Среднее арифметическое:

$$\bar{Q}_m = \tfrac{1}{m}\varepsilon_m + \tfrac{1}{m}\varepsilon_{m-1} + \tfrac{1}{m}\varepsilon_{m-2} + \ldots$$

$$\bar{Q}_m = \tfrac{1}{m}\varepsilon_m + (1 - \tfrac{1}{m})\bar{Q}_{m-1}$$

Экспоненциальное скользящее среднее (ЭСС):

$$\bar{Q}_m = \lambda\varepsilon_m + (1-\lambda)\lambda\varepsilon_{m-1} + (1-\lambda)^2\lambda\varepsilon_{m-2} + \ldots$$

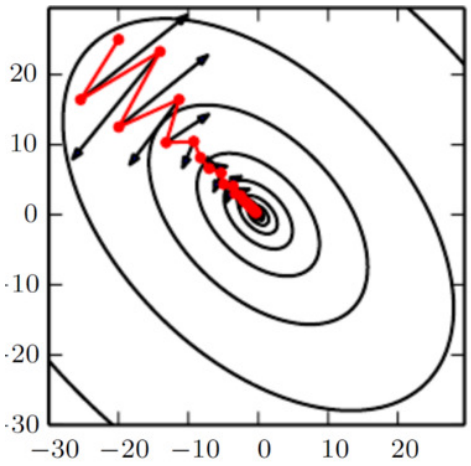
$$\bar{Q}_m = \lambda\varepsilon_m + (1 - \lambda)\bar{Q}_{m-1}$$

Параметр λ (порядка $\frac{1}{m}$) — *темп забывания* предыстории ряда.

Метод накопления инерции (momentum)

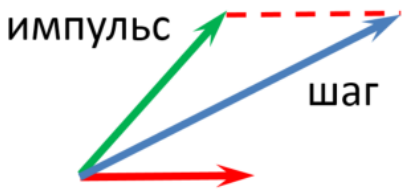
Momentum — экспоненциальное скользящее среднее градиента по последним $\approx \frac{1}{1-\gamma}$ итерациям [Б.Т.Поляк, 1964]:

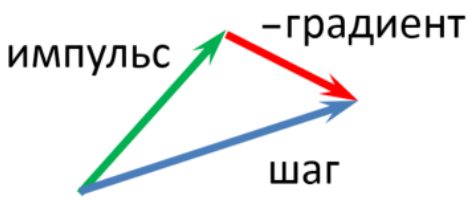
$$v := \gamma v + (1-\gamma)\nabla \mathscr{L}(w, x_i)$$
$$w := w - hv$$



NAG (Nesterov's accelerated gradient) — стохастический градиент с инерцией [Ю.Е.Нестеров, 1983]:

$$v := \gamma v + (1-\gamma)\nabla \mathscr{L}(w - h\gamma v, x_i)$$
$$w := w - hv$$





Эвристики

1. инициализация параметров w:
   случайная, по корреляции, комбинированная, ...
2. порядок предъявления объектов:
   чем больше ε_i, тем выше вероятность выбрать x_i
3. выбор градиентного шага: *метод скорейшего спуска*
   основан на поиске оптимального *адаптивного шага* h^*:
   $$\mathscr{L}\big(w - h\nabla\mathscr{L}(w, x_i), x_i\big) \to \min_h$$
4. метод Ньютона-Рафсона (второго порядка сходимости):
   $$w := w - h\big(\mathscr{L}''(w, x_i)\big)^{-1}\nabla\mathscr{L}(w, x_i), \quad \mathscr{L}'' \text{ — матрица Гессе}$$
   (в частности, диагональный метод Левенберга-Марквардта)
5. мультистарт: многократные запуски из разных случайных начальных приближений и выбор лучшего решения

Метод SG: Достоинства и недостатки

Достоинства:

1. *простота:* относительно легко реализуется
2. *универсальность:* для любых $a(x, w)$, $\mathscr{L}(w, x)$, $\mathscr{R}(w)$
3. *поточность:* возможность обучения на потоке данных
4. *подходит для обработки больших данных:*
   — можно получить неплохое решение, успев обработать лишь малую часть объектов
   — часто оказывается быстрее и лучше более сложных и ресурсоёмких методов второго порядка

Недостатки:

1. подбор комплекса эвристик является искусством (не забыть про переобучение, застревание, расходимость)

Оптимизационная задача обучения модели регрессии

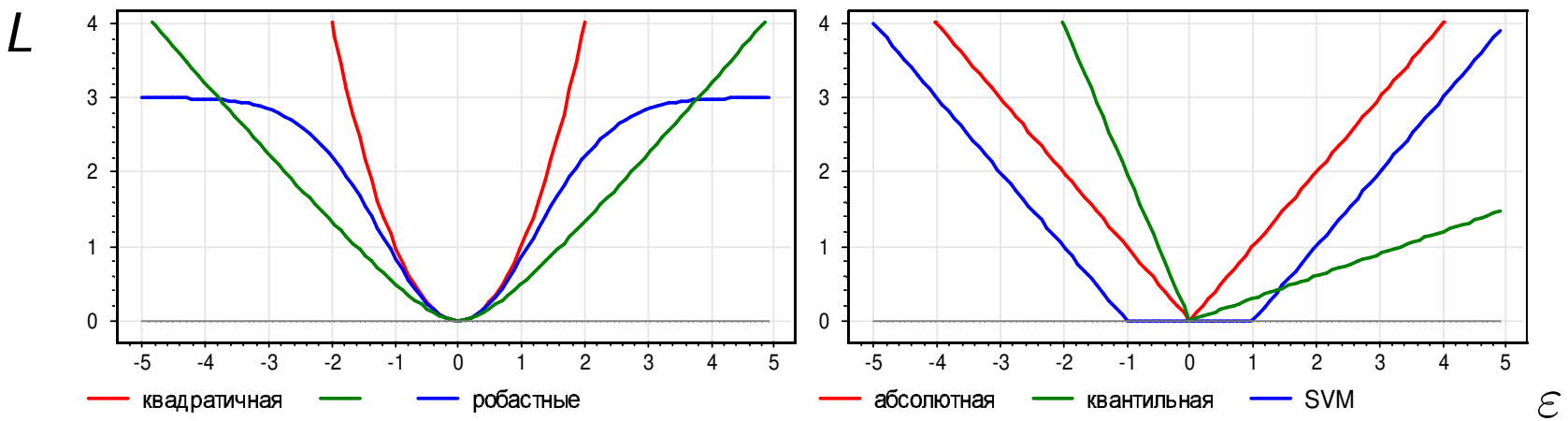
Дано: обучающая выборка $(x_i, y_i)_{i=1}^{\ell}$ с ответами $y_i \in \mathbb{R}$

Найти: вектор параметров w модели регрессии $a(x, w)$

Критерий: минимум эмпирического риска

$$\sum_{i=1}^{\ell} L\big(a(x_i, w) - y_i\big) \;\to\; \min_w$$

Унимодальные функции потерь $L(\varepsilon)$ от невязки $\varepsilon = a(x, w) - y$:



Пример. Задача прогнозирования объёмов продаж

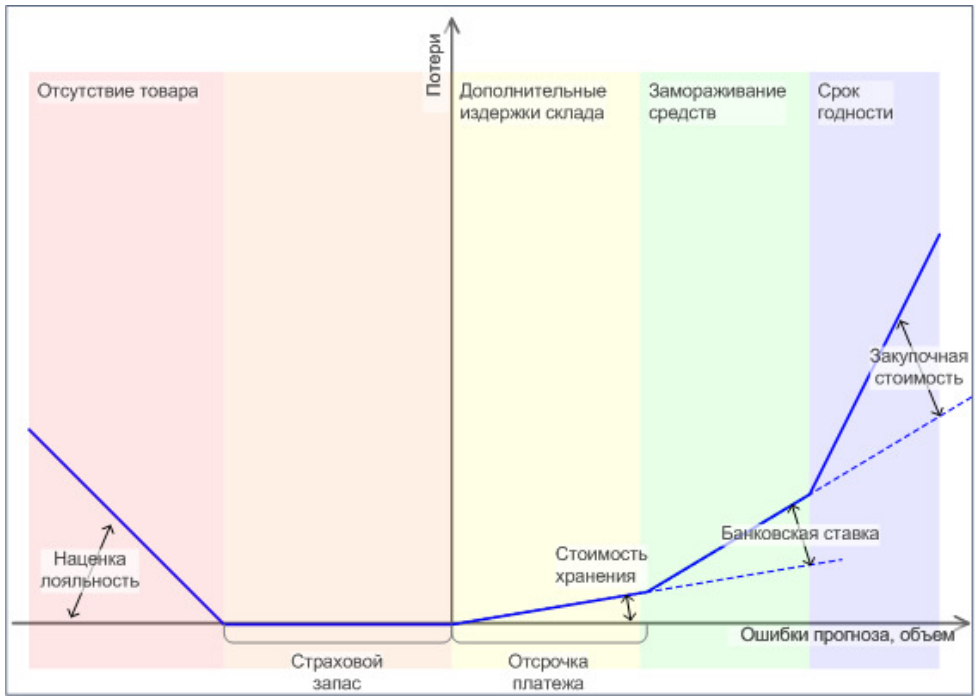
Объект — тройка $\langle$товар, магазин, день$\rangle$.

Примеры признаков:

- **бинарные:** выходной день, праздник, промоакция, и т. д.
- **количественные:** объёмы продаж в предшествующие дни.

Особенности задачи:

- функция потерь не квадратична и даже не симметрична;
- разреженные данные.



Вероятностная постановка задачи обучения регрессии

Модель данных с некоррелированным гауссовским шумом:

$$y_i = a(x_i, w) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2), \quad i = 1, \ldots, \ell$$

Метод максимума правдоподобия (ММП, MLE): при каком w плотность совместного распределения данных максимальна:

$$p(\varepsilon_1, \ldots, \varepsilon_\ell) = \prod_{i=1}^{\ell} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_i^2}\varepsilon_i^2\right) \to \max_w$$

$$-\ln p(\varepsilon_1, \ldots, \varepsilon_\ell) = C + \frac{1}{2}\sum_{i=1}^{\ell} \frac{1}{\sigma_i^2}\big(a(x_i, w) - y_i\big)^2 \to \min_w$$

- Постановки задачи МНК и ММП, совпадают, причём веса объектов связаны с дисперсией шума, $w_i = \sigma_i^{-2}$
- Использование неквадратичных функций потерь равносильно гипотезе о негауссовском распределение шума

Задача обучения модели двухклассовой классификации

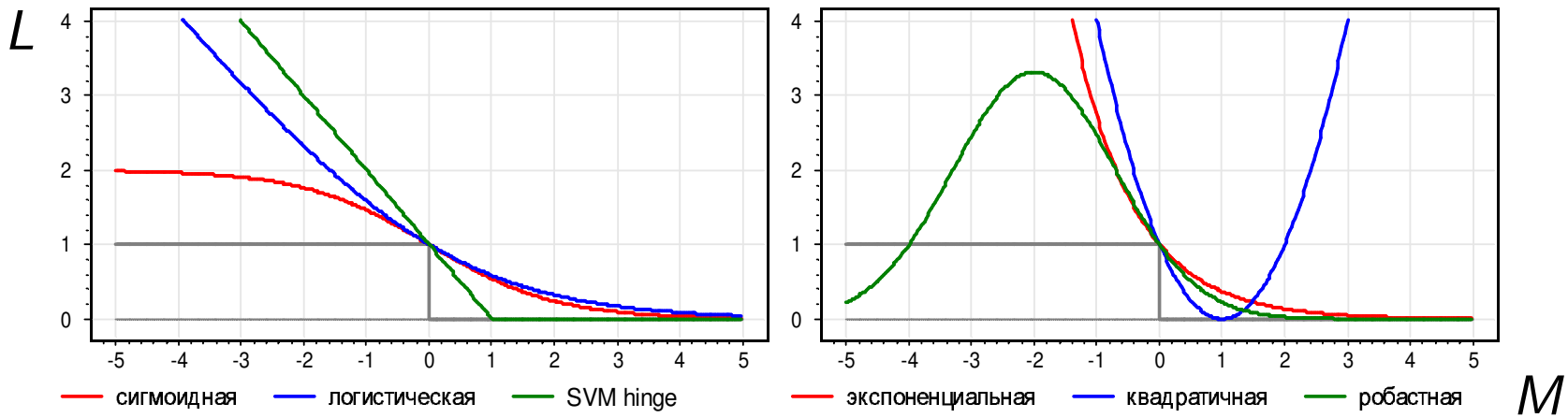
Дано: обучающая выборка $(x_i, y_i)_{i=1}^{\ell}$, $y_i \in \{-1, +1\}$

Найти: вектор w модели классификации $a(x, w) = \operatorname{sign} g(x, w)$

Критерий: min *аппроксимированного эмпирического риска*

$$\sum_{i=1}^{\ell} \big[g(x_i, w)y_i < 0\big] \leqslant \sum_{i=1}^{\ell} L\big(g(x_i, w)y_i\big) \;\to\; \min_w$$

Убывающие функции потерь $L(M)$ от *отступа* $M = g(x, w)y$:



Бинарный разделяющий классификатор (margin-based classifier)

Бинарный классификатор: $a(x, w) = \operatorname{sign} g(x, w), \quad Y = \{-1, +1\}$

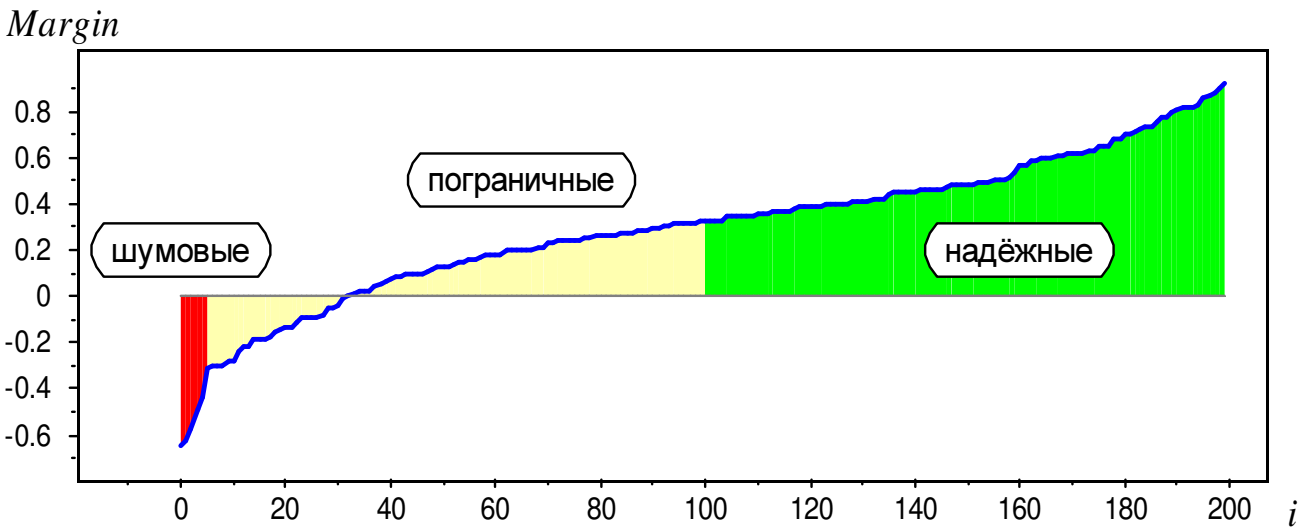
$g(x, w)$ — *разделяющая (дискриминантная) функция*

$x : g(x, w) = 0$ — разделяющая поверхность между классами

$M_i(w) = g(x_i, w)y_i$ — *отступ* (margin) объекта x_i

$M_i(w) < 0 \iff$ модель $a(x, w)$ ошибается на x_i

Ранжирование объектов по возрастанию отступов $M_i(w)$:



Задача многоклассовой классификации (multiclass classification)

Дано: обучающая выборка $(x_i, y_i)_{i=1}^{\ell}$, $y_i \in Y$, $|Y| < \infty$

Найти: вектор $w = (w_y : y \in Y)$ модели классификации

$$a(x, w) = \arg\max_{y \in Y} g_y(x, w_y)$$

Критерий: min *аппроксимированного эмпирического риска*

$$Q(w) = \sum_{i=1}^{\ell} \sum_{y \neq y_i} \left[g_{y_i}(x_i, w_{y_i}) < g_y(x_i, w_y) \right] \leqslant$$

$$\leqslant \sum_{i=1}^{\ell} \sum_{y \neq y_i} L\underbrace{\left(g_{y_i}(x_i, w_{y_i}) - g_y(x_i, w_y) \right)}_{M_{iy}(w)} \to \min_w$$

$L(M)$ — убывающая функция *отступа* $M_{iy}(w)$ по классу y

$M_i(w) = \min_{y \neq y_i} M_{iy}(w)$ — *отступ* (margin) объекта x_i

$M_i(w) < 0 \iff$ модель $a(x, w)$ ошибается на x_i

Вероятностная постановка задачи классификации

Дано: простая (i.i.d.) выборка $(x_i, y_i)_{i=1}^{\ell} \sim p(x, y)$, порождаемая в.п. $X \times Y$ с неизвестной плотностью $p(x, y)$

Найти: модель плотности $p(x, y; w) = P(y|x, w)p(x)$, где $P(y|x, w)$ — модель условной вероятности класса с параметром w
$p(x)$ — неизвестное и непараметризуемое распределение на X

Критерий: максимум правдоподобия, т.е. max плотности совместного распределения всей выборки данных:

$$\prod_{i=1}^{\ell} p(x_i, y_i; w) = \prod_{i=1}^{\ell} P(y_i|x_i, w)\cancel{p(x_i)} \to \max_w$$

Логарифм правдоподобия (log-likelihood, log-loss):

$$Q(w) = -\sum_{i=1}^{\ell} \log P(y_i|x_i, w) \to \min_w$$

Связь правдоподобия и аппроксимации эмпирического риска

Максимизация логарифма правдоподобия,
$P(y|x,w)$ — модель *условной вероятности класса*:

$$Q(w) = -\sum_{i=1}^{\ell} \log P(y_i|x_i,w) \to \min_w$$

Минимизация аппроксимированного эмпирического риска,
$g(x,w)$ — модель *разделяющей поверхности*, $Y = \{\pm 1\}$:

$$Q(w) = \sum_{i=1}^{\ell} \mathscr{L}\big(y_i g(x_i,w)\big) \to \min_w;$$

Эти два принципа эквивалентны, если положить

$$-\log P(y_i|x_i,w) = \mathscr{L}\big(y_i g(x_i,w)\big).$$

$$\boxed{\text{модель } P(y|x,w)} \rightleftarrows \boxed{\text{модель } g(x,w) \text{ и } \mathscr{L}(M)}.$$

Анализ ошибок классификации

Задача классификации на два класса: y_i, $a(x_i) \in \{-1, +1\}$.

	модель классификации	учитель
TP, True Positive	$a(x_i) = +1$	$y_i = +1$
TN, True Negative	$a(x_i) = -1$	$y_i = -1$
FP, False Positive	$a(x_i) = +1$	$y_i = -1$
FN, False Negative	$a(x_i) = -1$	$y_i = +1$

FP: ложноположительно, ошибка I рода, «ложная тревога»

FN: ложноотрицательно, ошибка II рода, «пропуск цели»

Правильность классификации (чем больше, тем лучше):

$$\text{Accuracy} = \frac{1}{\ell}\sum_{i=1}^{\ell}[a(x_i) = y_i] = \frac{\text{TP} + \text{TN}}{\text{FP} + \text{FN} + \text{TP} + \text{TN}}$$

Недостаток: не учитывается дисбаланс численности классов, а также различие цены ошибки I и II рода.

Определение ROC-кривой (Receiver Operating Characteristic)

Введём порог w_0 в модель: $a(x; w, w_0) = \text{sign}\big(g(x, w) - w_0\big)$
(чем больше w_0, тем больше x_i, на которых $a(x_i) = -1$)

- по оси X: доля *ошибочных положительных классификаций* (FPR — false positive rate):

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} = \frac{\sum_{i=1}^{\ell}[y_i = -1][a(x_i; w, w_0) = +1]}{\sum_{i=1}^{\ell}[y_i = -1]}$$

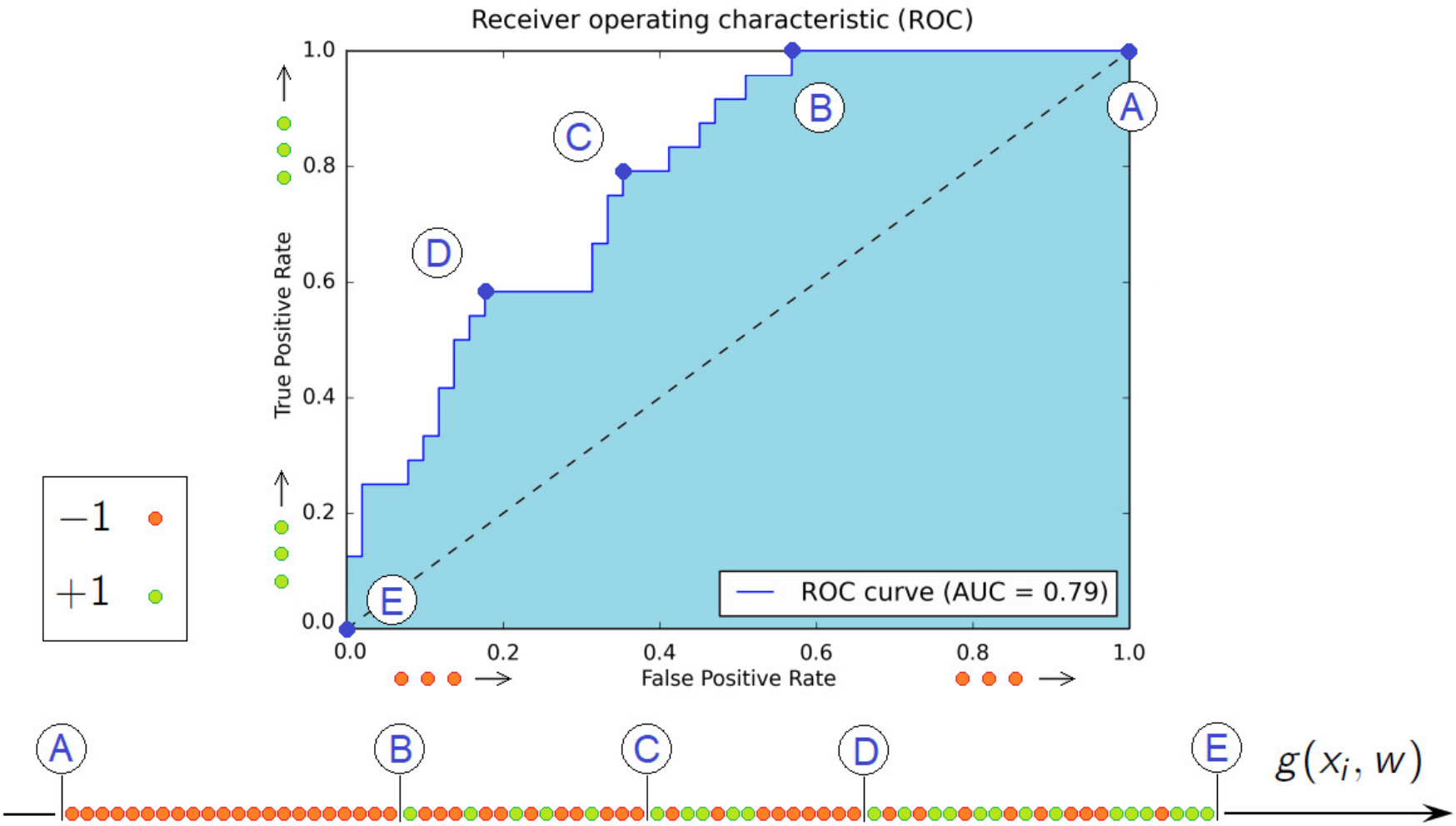
  $1 - \text{FPR}$ называется *специфичностью* алгоритма a

- по оси Y: доля *правильных положительных классификаций* (TPR — true positive rate):

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\sum_{i=1}^{\ell}[y_i = +1][a(x_i; w, w_0) = +1]}{\sum_{i=1}^{\ell}[y_i = +1]}$$

  TPR называется также *чувствительностью* алгоритма a

ROC-кривая и площадь под кривой AUC (Area Under Curve)



ABCDE — положения порога w_0 на оси значений функции g

Задача максимизации площади под кривой ROC-AUC

Модель классификации: $a(x_i, w, w_0) = \text{sign}\big(g(x_i, w) - w_0\big)$

AUC — это доля правильно упорядоченных пар (x_i, x_j):

$$\text{AUC}(w) = \frac{1}{\ell_-}\sum_{i=1}^{\ell}\big[y_i = -1\big]\text{TPR}_i =$$

$$= \frac{1}{\ell_-\ell_+}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}\big[y_i < y_j\big]\big[g(x_i, w) < g(x_j, w)\big] \to \max_w$$

Критерий: максимум аппроксимированного AUC:

$$1 - \text{AUC}(w) \leqslant Q(w) = \sum_{i,j:\ y_i<y_j} L\underbrace{\big(g(x_j, w) - g(x_i, w)\big)}_{M_{ij}(w)} \to \min_w$$

где $L(M)$ — убывающая функция попарного отступа $M_{ij}(w)$

SG: градиентные шаги по парам объектов (x_i, x_j): $y_i < y_j$

Задача обучения ранжированию

Дано: обучающая выборка $(x_1, \ldots, x_\ell)$,
$i \prec j$ — отношение «x_j лучше, чем x_i» между объектами из X^ℓ

Найти: параметры w модели ранжирования $a(x, w)$,
восстанавливающей правильное отношение порядка:

$$i \prec j \quad \Rightarrow \quad a(x_i, w) < a(x_j, w)$$

Критерий: число неверно ранжированных пар объектов

$$Q(w) = \sum_{i \prec j} \big[\underbrace{a(x_j, w) - a(x_i, w)}_{M_{ij}(w)} < 0 \big]$$

$$\leqslant \sum_{i \prec j} L\big(a(x_j, w) - a(x_i, w)\big) \;\to\; \min_w$$

где $L(M)$ — убывающая функция попарного отступа $M_{ij}(w)$
SG: градиентные шаги по парам объектов (x_i, x_j): $i \prec j$
Максимизация AUC сводится к бинарному ранжированию!

Примеры задач ранжирования

Ранжирование (Learning to Rank, LtR, L2R, LETOR) нужно в системах человеко-машинного принятия решений, когда пользователь выбирает из ранжированного списка вариантов

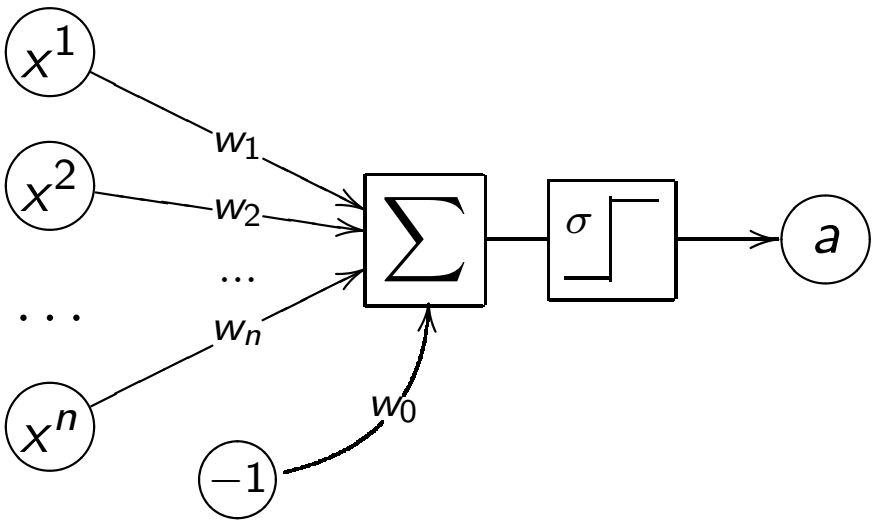
- ранжирование выдачи поисковой системы
- ранжирование рекомендаций пользователям (книги, фильмы, музыка, товары интернет-магазина, и т.п.)
- ранжирование вариантов автоматического завершения запроса (Query Auto Completion, auto-suggest)
- ранжирование возможных ответов в диалоговых системах (Question Answering Systems)
- ранжирование вариантов перевода в системах машинного перевода (Machine Translation)

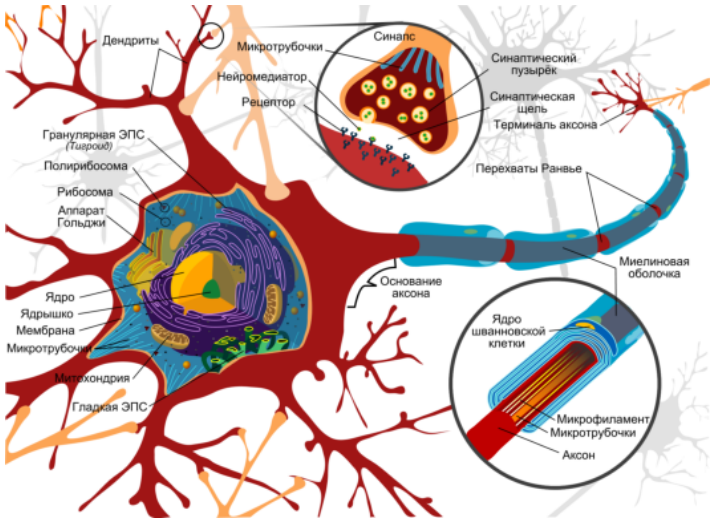
Линейный классификатор — математическая модель нейрона

Линейная модель нейрона МакКаллока-Питтса [1943]:

$$a(x,w) = \sigma\left(\sum_{j=1}^{n} w_j f_j(x) - w_0\right) = \sigma\big(\langle w,x\rangle - w_0\big),$$

$\sigma(z)$ — функция активации (например, sign, th или $\frac{1}{1+e^{-z}}$),
w_j — весовые коэффициенты синаптических связей,
w_0 — порог активации





Двухклассовая логистическая регрессия

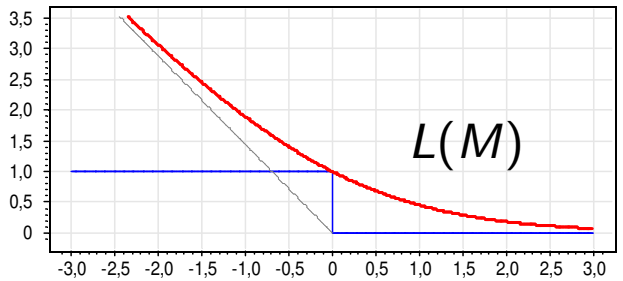
Линейная модель классификации для двух классов $Y = \{-1, 1\}$:

$$a(x, w) = \text{sign}\langle w, x\rangle, \quad x, w \in \mathbb{R}^n$$

Отступ $M = \langle w, x\rangle y$

Логарифмическая функция потерь:

$$L(M) = \log\left(1 + e^{-M}\right)$$

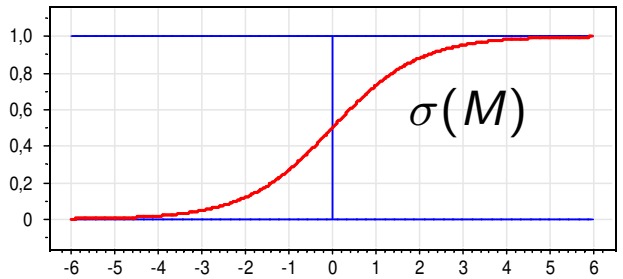


Модель условной вероятности:

$$P(y|x, w) = \sigma(M) = \frac{1}{1+e^{-M}},$$

где $\sigma(M)$ — сигмоидная функция,

важное свойство: $\sigma(-M) = 1 - \sigma(M)$



Максимизация правдоподобия (log-loss):

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

Многоклассовая логистическая регрессия

Линейный классификатор при произвольном числе классов $|Y|$:

$$a(x) = \arg\max_{y \in Y} \langle w_y, x \rangle, \quad x, w_y \in \mathbb{R}^n$$

Вероятность того, что объект x относится к классу y:

$$P(y|x, w) = \frac{\exp\langle w_y, x \rangle}{\sum_{z \in Y} \exp\langle w_z, x \rangle} = \underset{y \in Y}{\text{SoftMax}} \langle w_y, x \rangle,$$

функция SoftMax: $\mathbb{R}^Y \to \mathbb{R}^Y$ переводит произвольный вектор в нормированный вектор дискретного распределения.

Максимизация правдоподобия (log-loss):

$$Q(w) = -\sum_{i=1}^{\ell} \log P(y_i|x_i, w) \to \min_w$$

Мультиколлинеарность и переобучение в линейных моделях

Возможные причины переобучения:

- слишком мало объектов, слишком много признаков
- линейная зависимость (мультиколлинеарность) признаков:

  пусть построен классификатор: $a(x, w) = \text{sign}\langle w, x\rangle$
  мультиколлинеарность: $\exists u \in \mathbb{R}^n$: $\forall x \in X \;\; \langle u, x\rangle = 0$
  неединственность решения: $\forall \gamma \in \mathbb{R} \;\; a(x, w) = \text{sign}\langle w + \gamma u, x\rangle$

Проявления переобучения:

- слишком большие веса $|w_j|$ разных знаков
- неустойчивость дискриминантной функции $\langle w, x\rangle$
- $Q(X^\ell) \ll Q(X^k)$

Простой способ уменьшить переобучение:

- регуляризация $\|w\| \to \min$ (сокращение весов, weight decay)

Регуляризация (сокращение весов, weight decay)

Штраф за увеличение нормы вектора весов:

$$\mathscr{L}_\tau(w, x_i) = \mathscr{L}(w, x_i) + \frac{\tau}{2}\|w\|^2 = \mathscr{L}(w, x_i) + \frac{\tau}{2}\sum_{j=1}^{n} w_j^2 \to \min_w .$$

Градиент:

$$\nabla\mathscr{L}_\tau(w, x_i) = \nabla\mathscr{L}(w, x_i) + \tau w.$$

Модификация градиентного шага в методе SG:

$$w := w(1 - h\tau) - h\nabla\mathscr{L}(w, x_i).$$

Методы подбора коэффициента регуляризации τ:

1. hold-out или скользящий контроль
2. стохастическая адаптация

Вероятностный смысл регуляризации

$P(y|x,w)$ — вероятностная модель данных;
$p(w;\gamma)$ — априорное распределение параметров модели;
γ — вектор *гиперпараметров*;

Теперь не только появление выборки X^ℓ,
но и появление модели w также полагается стохастическим.

Совместное правдоподобие данных и модели:

$$p(X^\ell, w) = p(X^\ell|w)\, p(w;\gamma).$$

Принцип максимума апостериорной вероятности
(Maximum a Posteriori Probability, MAP):

$$Q_{\text{MAP}}(w) = \ln p(X^\ell, w) = \underbrace{\sum_{i=1}^{\ell} \log P(y_i|x_i,w)}_{Q_{\text{MLE}}(w)} + \underbrace{\log p(w;\gamma)}_{\text{регуляризатор}} \to \max_w$$

Примеры: априорные распределения Гаусса и Лапласа

Пусть веса w_j независимы, $\mathsf{E}w_j = 0$, $\mathsf{D}w_j = C$.

Распределение Гаусса и квадратичный (L_2) регуляризатор:

$$p(w;C) = \frac{1}{(2\pi C)^{n/2}} \exp\left(-\frac{\|w\|^2}{2C}\right), \quad \|w\|^2 = \sum_{j=1}^{n} w_j^2,$$

$$-\ln p(w;C) = \tfrac{1}{2C}\|w\|^2 + \cancel{\text{const}}$$

Распределение Лапласа и абсолютный (L_1) регуляризатор:

$$p(w;C) = \frac{1}{(2C)^{n}} \exp\left(-\frac{\|w\|}{C}\right), \quad \|w\| = \sum_{j=1}^{n} |w_j|,$$

$$-\ln p(w;C) = \tfrac{1}{C}\|w\| + \cancel{\text{const}}$$

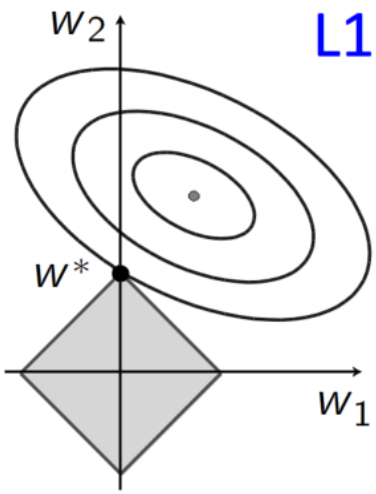
C — гиперпараметр, $\tau = \frac{1}{C}$ — коэффициент регуляризации.

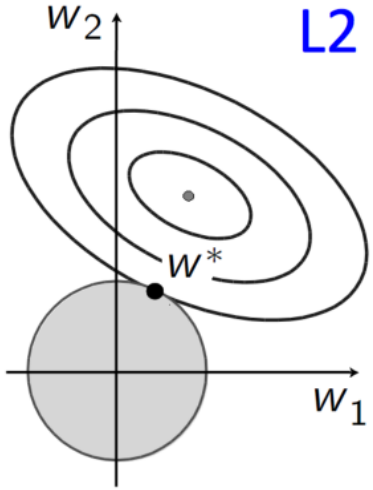
Регуляризация по L_1-норме для отбора признаков

LASSO — Least Absolute Shrinkage and Selection Operator

$$Q(w) + \tau \sum_{j=1}^{n} |w_j| \to \min_w \quad \Longleftrightarrow \quad \begin{cases} Q(w) \to \min\limits_w; \\ \sum\limits_{j=1}^{n} |w_j| \leqslant \varkappa; \end{cases}$$

$$\sum_{j=1}^{n} |w_j| \leqslant \varkappa$$



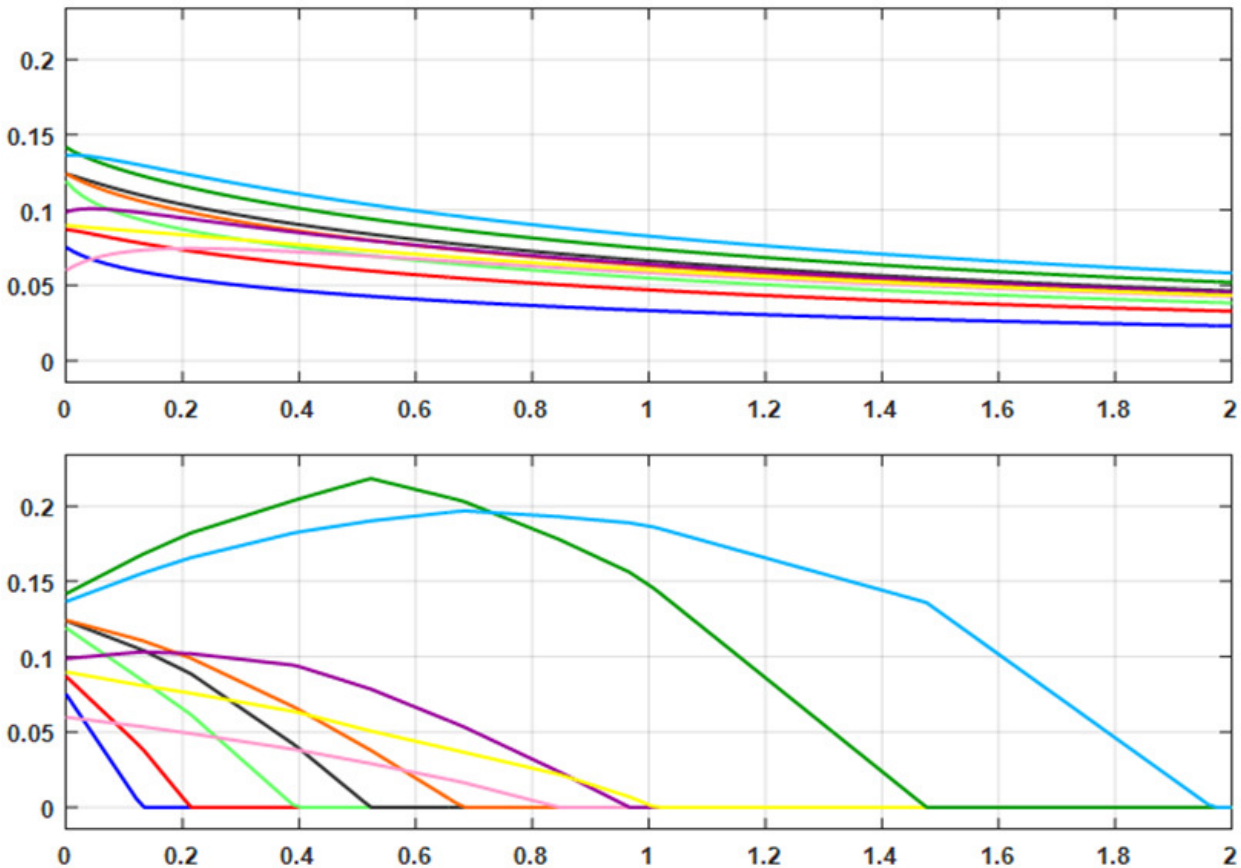


$$\sum_{j=1}^{n} w_j^2 \leqslant \varkappa$$

T.Hastie, R.Tibshirani, J.Friedman. The Elements of Statistical Learning. 2017.

Сравнение L_2 и L_1 регуляризации

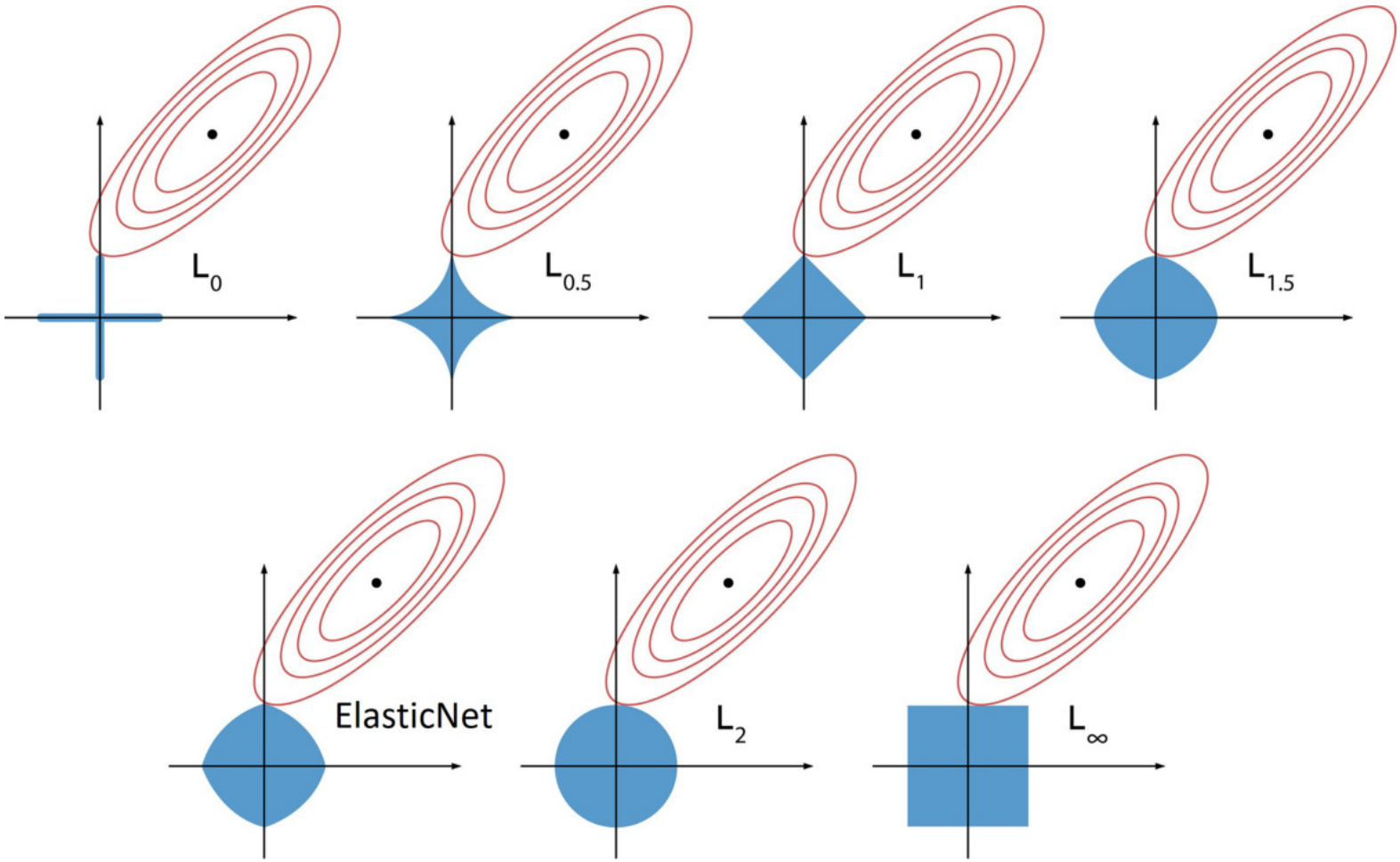
Зависимость весов w_j от параметра селективности τ



В LASSO с увеличением τ усиливается отбор признаков

Геометрическая интерпретация отбора признаков

Сравнение регуляризаторов по различным L_p-нормам:



Пример. Бинаризация признаков и скоринговая карта

Задача кредитного скоринга:

- x_i — заёмщики
- $y_i = -1\,(\text{bad}), +1\,(\text{good})$

Бинаризация признаков $f_j(x)$:

$$b_{jk}(x) = \big[f_j(x) \text{ из } k\text{-го интервала}\big]$$

Линейная модель классификации:

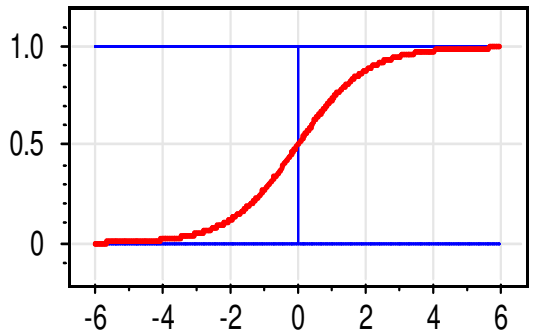
$$a(x, w) = \text{sign} \sum_{j,k} w_{jk}\, b_{jk}(x).$$

Вес признака w_{jk} равен его вкладу в общую сумму баллов (score).

признак j	интервал k	w_{jk}
Возраст	до 25	5
	25 - 40	10
	40 - 50	15
	50 и больше	10
Собственность	владелец	20
	совладелец	15
	съемщик	10
	другое	5
Работа	руководитель	15
	менеджер среднего звена	10
	служащий	5
	другое	0
Стаж	1/безработный	0
	1..3	5
	3..10	10
	10 и больше	15
Работа_мужа /жены	нет/домохозяйка	0
	руководитель	10
	менеджер среднего звена	5
	служащий	1

Оценивание рисков в скоринге

Логистическая регрессия не только определяет веса w, но и оценивает *апостериорные вероятности* классов



$$\mathrm{P}(y|x,w) = \frac{1}{1+e^{-\langle w,x\rangle y}}$$

Оценка *риска* (математического ожидания) потерь объекта x:

$$R(x) = \sum_{y\in Y} D_{xy}\,\mathrm{P}(y|x,w),$$

где D_{xy} — величина потери для объекта x с исходом y.

Оценка говорит о том, сколько мы потеряем в среднем.
Но сколько мы потеряем в худшем случае?

Методика VaR (Value at Risk)

Стохастическое моделирование: $N = 10^4$ раз

- для каждого x_i разыгрывается исход $y_i \sim P(y|x_i)$;
- вычисляется сумма потерь по портфелю $V = \sum_{i=1}^{\ell} D_{x_i y_i}$;

99%-квантиль эмпирического распределения потерь определяет величину резервируемого капитала

Резюме в конце лекции

- Метод стохастического градиента (SG)
  — подходит для любых моделей и функций потерь
  — подходит для обучения по большим данным
- *Аппроксимация пороговой функции потерь* $L(M)$
  — общий приём для классификации, ранжирования
  — позволяет использовать градиентную оптимизацию в задачах с исходно дискретнозначными критериями
- *Регуляризация* снижает переобучение, возникающее в линейных моделях из-за мультиколлинеарности
- *Логистическая регрессия* — метод классификации, оценивающий условные вероятности классов $P(y|x)$
- *AUC* — мера качества классификации, не зависящая от соотношения штрафов и численности классов