

- Математические модели машинного обучения •

Метод стохастического градиента и линейные модели классификации

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`

«Машинное обучение (курс лекций, К.В.Воронцов)»

1 Градиентная оптимизация в машинном обучении

- Оптимизационные задачи в машинном обучении
- Метод стохастического градиента
- Ускорение сходимости и другие эвристики

2 Модели классификации на основе разделимости

- Разделяющие модели классификации
- Анализ ошибок классификации
- Линейные модели, переобучение и регуляризация

3 Вероятностные функции потерь

- Логистическая регрессия
- Вероятностная модель классификации
- Пример. Задача кредитного скоринга

Общая постановка задачи обучения по прецедентам

Дано: X — пространство объектов

$X^\ell = \{x_1, \dots, x_\ell\} \subset X$ — обучающая выборка, прецеденты

$a(x, w)$, $a: X \times W \rightarrow Y$ — параметрическая модель, гипотеза

Найти $w \in W$ — вектор параметров модели $a(x, w)$

Критерий Empirical Risk Minimization (ERM) — минимизация эмпирического риска, в общем случае с регуляризацией:

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

$\mathcal{L}(w, x)$ — функция потерь (loss function),

тем больше, чем хуже ответ модели $a(x, w)$ на объекте x

$\mathcal{R}(w)$ — регуляризатор, априорные требования к модели

τ — коэффициент регуляризации для балансировки критериев

Градиентный метод минимизации эмпирического риска

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

Метод градиентного спуска:

$w^{(0)}$:= начальное приближение;

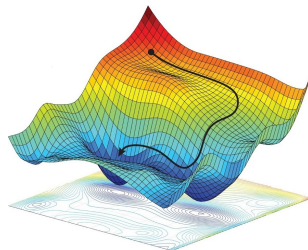
$$w^{(t+1)} := w^{(t)} - h \nabla_w Q(w^{(t)})$$

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

h — градиентный шаг, называемый также темпом обучения

$$w^{(t+1)} := w^{(t)} - h \left(\frac{1}{\ell} \sum_{i=1}^{\ell} \nabla_w \mathcal{L}(w^{(t)}, x_i) + \tau \nabla_w \mathcal{R}(w^{(t)}) \right)$$

Идея ускорения сходимости: брать случайное подмножество слагаемых, или даже один объект, сразу обновляя вектор весов



Метод стохастического градиента SG (Stochastic Gradient)

$$\text{Критерий ERM: } Q(w) = \frac{1}{\ell} \sum_{i=1}^{\ell} \mathcal{L}(w, x_i) + \tau \mathcal{R}(w) \rightarrow \min_w$$

Вход: выборка X^ℓ , параметры h, τ, λ ;

Выход: вектор весов w ;

- 1 инициализировать веса $w_j, j = 1, \dots, N$;
- 2 инициализировать оценку $Q(w)$ по небольшой подвыборке;
- 3 **повторять**
 - 4 объект x_i выбрать из X^ℓ случайным образом;
 - 5 потеря: $\mathcal{L}_i := \mathcal{L}(w, x_i)$;
 - 6 градиентный шаг: $w := w - h \nabla \mathcal{L}(w, x_i) - h\tau \nabla \mathcal{R}(w)$;
 - 7 рекуррентная оценка критерия: $Q := \lambda \mathcal{L}_i + (1 - \lambda)Q$;
- 8 **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 $x_1, \dots, x_\ell$ намного дольше градиентного шага по одному объекту x_i .

Решение: использовать приближённую рекуррентную формулу.

Среднее арифметическое:

$$\bar{Q}_m = \frac{1}{m} \mathcal{L}_m + \frac{1}{m} \mathcal{L}_{m-1} + \frac{1}{m} \mathcal{L}_{m-2} + \dots$$

$$\bar{Q}_m = \frac{1}{m} \mathcal{L}_m + (1 - \frac{1}{m}) \bar{Q}_{m-1}$$

Экспоненциальное скользящее среднее (ЭСС):

$$\bar{Q}_m = \lambda \mathcal{L}_m + (1 - \lambda) \lambda \mathcal{L}_{m-1} + (1 - \lambda)^2 \lambda \mathcal{L}_{m-2} + \dots$$

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

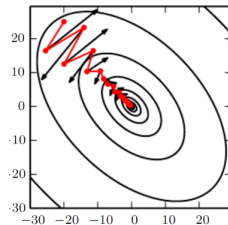
Параметр λ (порядка $\frac{1}{m}$) — *темп забывания* предыстории ряда.

Метод накопления инерции (momentum)

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

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

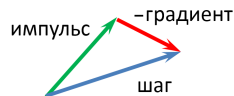
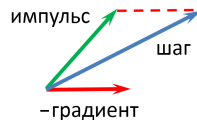
$$w := w - hv$$



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

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

$$w := w - hv$$



Эвристики

- 1 инициализация параметров w :
случайная, по корреляции, комбинированная, ...
- 2 порядок предъявления объектов:
чем больше $\mathcal{L}_i$, тем выше вероятность выбрать x_i
- 3 выбор градиентного шага: *метод скорейшего спуска*
основан на поиске оптимального *адаптивного шага* h^* :

$$\mathcal{L}(w - h\nabla\mathcal{L}(w, x_i), x_i) \rightarrow \min_h$$

- 4 мультистарт: многократные запуски из разных случайных начальных приближений и выбор лучшего решения
- 5 метод Ньютона-Рафсона (второго порядка сходимости):

$$w := w - h(\mathcal{L}''(w, x_i))^{-1}\nabla\mathcal{L}(w, x_i), \quad \mathcal{L}'' \text{ — матрица Гессе}$$

(в частности, диагональный метод Левенберга-Марквардта)

Диагональный метод Левенберга-Марквардта

Метод Ньютона-Рафсона:

$$w := w - h (\mathcal{L}''(w, x_i))^{-1} \nabla \mathcal{L}(w, x_i),$$

где $\mathcal{L}''(w, x_i) = \left(\frac{\partial^2 \mathcal{L}(w, x_i)}{\partial w_j \partial w_{j'}} \right)$ — гессиан, $n \times n$ -матрица

Эвристика. Считаем, что гессиан диагонален:

$$w_j := w_j - h \max \left\{ \mu, \frac{\partial^2 \mathcal{L}(w, x_i)}{\partial w_j^2} \right\}^{-1} \frac{\partial \mathcal{L}(w, x_i)}{\partial w_j},$$

$h > 0$ — темп обучения, можно полагать $h = 1$,

$\mu > 0$ — параметр, предотвращающий обнуление знаменателя.

Вблизи минимума сходимость второго порядка с темпом h

Вдали от минимума сходимость первого порядка с темпом $\frac{h}{\mu}$

Метод SG: достоинства и недостатки

Достоинства:

- 1 *простота*: относительно легко реализуется
- 2 *универсальность*: для любых $a(x, w)$, $\mathcal{L}(w, x)$, $\mathcal{R}(w)$
- 3 *поточность*: возможность обучения на потоке данных
- 4 *подходит для обработки больших данных*:
 - можно получить неплохое решение, успев обработать лишь малую часть обучающей выборки
 - 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)

Задача бинарной классификации

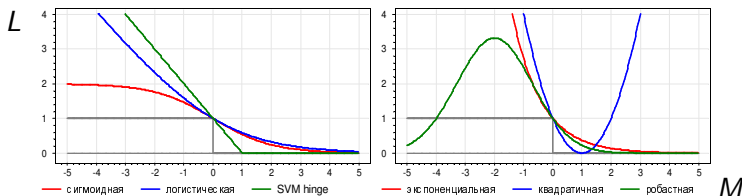
Дано: обучающая выборка $(x_i, y_i)_{i=1}^{\ell}$, $y_i \in \{-1, +1\}$

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

Критерий: $\min$ аппроксимированного эмпирического риска

$$\sum_{i=1}^{\ell} [b(x_i, w)y_i < 0] \leq \sum_{i=1}^{\ell} L(b(x_i, w)y_i) \rightarrow \min_w$$

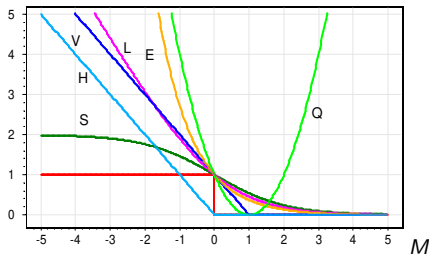
Убывающие функции потерь $L(M)$ от отступа $M = b(x, w)y(x)$:



Каждая из этих функций придаёт модели особые свойства

Непрерывные аппроксимации пороговой функции потерь

Выбор функции потерь $L(M)$ порождает большое разнообразие методов, отличающихся полезными свойствами (это спойлер)



$[M < 0]$

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

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

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

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

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

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

пороговая функция потерь

кусочно-линейная (метод опорных векторов, SVM)

кусочно-линейная (правило Хэбба, Hebb's rule)

логарифмическая (логистическая регрессия, LR)

квадратичная (линейный дискриминант Фишера, FLD)

сигмоидная (искусственные нейронные сети, ANN)

экспоненциальная (линейный ансамбль AdaBoost)

Бинарный разделяющий классификатор (margin-based classifier)

Бинарный классификатор: $a(x, w) = \text{sign } b(x, w)$, $Y = \{-1, +1\}$

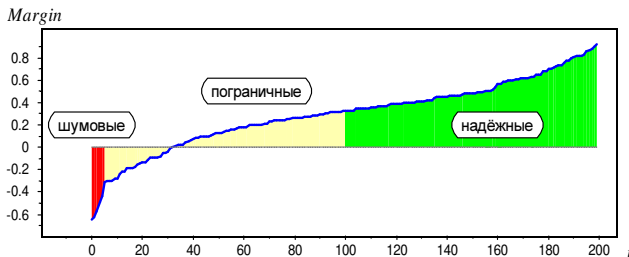
$b(x, w)$ — разделяющая (дискриминантная) функция

$\{x: b(x, w) = 0\}$ — разделяющая поверхность между классами

$M_i(w) = b(x_i, w)y_i$ — 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

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

где $b_y(x, w_y)$ отделяет объекты класса y от всех остальных,
 $M_{iy}(w) = b_{y_i}(x_i, w_{y_i}) - b_y(x_i, w_y)$ — отступ объекта x_i по классу y

Критерий ERM «каждый против всех» (one-vs.-all, OvA):

$$Q(w) = \sum_{i=1}^{\ell} \sum_{y \neq y_i} [M_{iy}(w) < 0] \leq \sum_{i=1}^{\ell} \sum_{y \neq y_i} L(M_{iy}(w)) \rightarrow \min_w,$$

где функция $L(M)$ выбирается, как правило, невозрастающей

Ошибки первого и второго рода

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

	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.

ROC-кривая (Receiver Operating Characteristic)

Бинарный классификатор: $a(x; w, w_0) = \text{sign}(b(x, w) - w_0)$

Кривая ROC: как меняется качество a при варьировании w_0
(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):

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

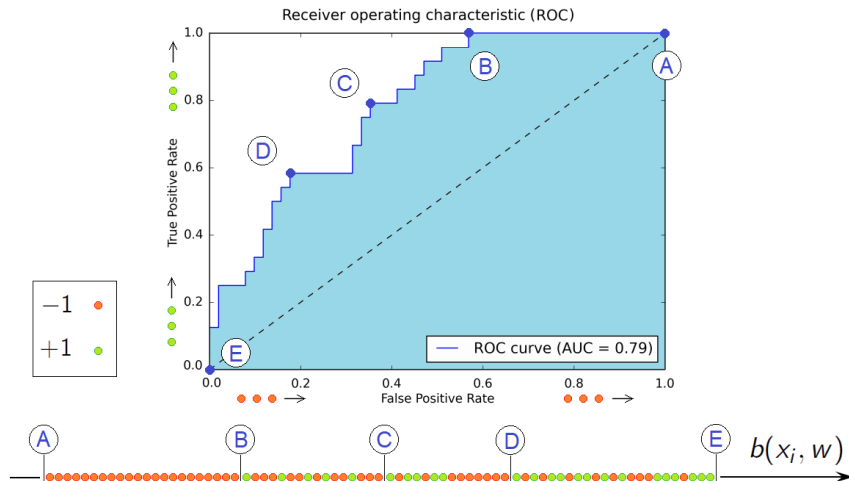
$1 - \text{FPR}(a)$ называется специфичностью алгоритма a

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

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

$\text{TPR}(a)$ называется также чувствительностью алгоритма a

Площадь под ROC-кривой AUC (Area Under Curve)



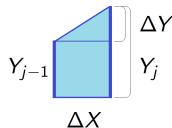
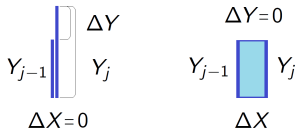
ABCDE — положения порога w_0 на оси значений функции b

Алгоритм эффективного построения ROC-кривой

Вход: выборка $\{x_i\}_{i=1}^{\ell}$; дискриминантная функция $b(x, w)$;

Выход: ROC-кривая $(X_j, Y_j)_{j=0}^k$, $k \leq \ell$ и площадь AUC

- 1 $\ell_y := \sum_{i=1}^{\ell} [y_i = y]$, для всех $y \in Y$;
- 2 упорядочить $\{x_i\}$ по убыванию $b_i = b(x_i, w)$: $b_1 \geq \dots \geq b_{\ell}$;
- 3 $(X_0, Y_0) := (0, 0)$; $AUC := 0$; $\Delta X := 0$; $\Delta Y := 0$; $j := 1$;
- 4 **для** $i := 1, \dots, \ell$
- 5 $\Delta X := \Delta X + \frac{1}{\ell_-} [y_i = -1]$;
- 6 $\Delta Y := \Delta Y + \frac{1}{\ell_+} [y_i = +1]$;
- 7 **если** $(b_i \neq b_{i-1})$ **то**
- 8 $X_j := X_{j-1} + \Delta X$;
- 9 $Y_j := Y_{j-1} + \Delta Y$;
- 10 $AUC := AUC + \frac{1}{2} (Y_{j-1} + Y_j) \Delta X$;
- 11 $j := j + 1$; $\Delta X := 0$; $\Delta Y := 0$;



Задача максимизации площади под кривой ROC-AUC

Бинарный классификатор: $a(x; w, w_0) = \text{sign}(b(x, w) - w_0)$

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

$$\begin{aligned} \text{AUC}(w) &= \frac{1}{\ell_-} \sum_{i=1}^{\ell} [y_i = -1] \text{TPR}_i = \\ &= \frac{1}{\ell_- \ell_+} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} [y_i < y_j] [b(x_i, w) < b(x_j, w)] \rightarrow \max_w \end{aligned}$$

Критерий: максимум аппроксимированного AUC:

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

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

Метод SG: шаг по случайным парам объектов (x_i, x_j) : $y_i < y_j$

$$w := w - h L'(M_{ij}(w)) \cdot \nabla_w (b(x_j, w) - b(x_i, w))$$

Точность и полнота бинарной классификации

В информационном поиске не важен TN:

$$\text{Точность, Precision} = \frac{TP}{TP+FP}$$

$$\text{Полнота, Recall} = \frac{TP}{TP+FN}$$

Precision — доля релевантных среди найденных

Recall — доля найденных среди релевантных

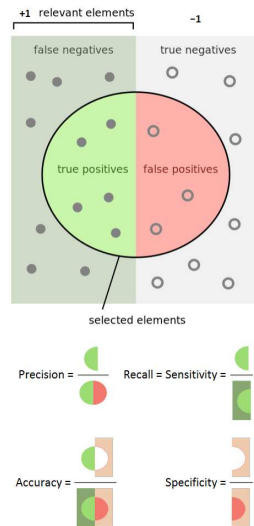
В медицинской диагностике важно всё:

$$\text{Чувствительность, Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Специфичность, Specificity} = \frac{TN}{TN+FP}$$

Sensitivity — доля верных положительных диагнозов

Specificity — доля верных отрицательных диагнозов



Точность и полнота многоклассовой классификации

Для каждого класса $y \in Y$:

TP_y — верные положительные

FP_y — ложные положительные

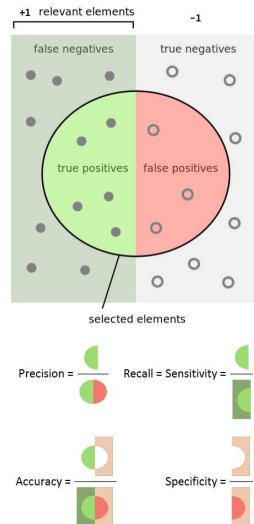
FN_y — ложные отрицательные

Точность и полнота с микроусреднением:

$$\text{Precision: } P = \frac{\sum_y TP_y}{\sum_y (TP_y + FP_y)};$$

$$\text{Recall: } R = \frac{\sum_y TP_y}{\sum_y (TP_y + FN_y)};$$

Микроусреднение не чувствительно
к ошибкам на малочисленных классах



Точность и полнота многоклассовой классификации

Для каждого класса $y \in Y$:

TP_y — верные положительные

FP_y — ложные положительные

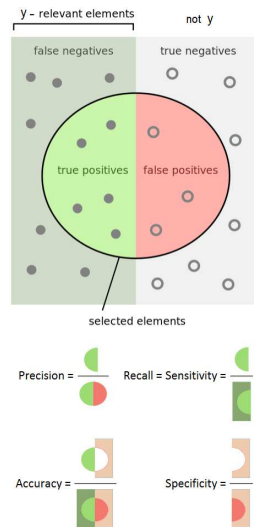
FN_y — ложные отрицательные

Точность и полнота с макроусреднением:

$$\text{Precision: } P = \frac{1}{|Y|} \sum_y \frac{TP_y}{TP_y + FP_y};$$

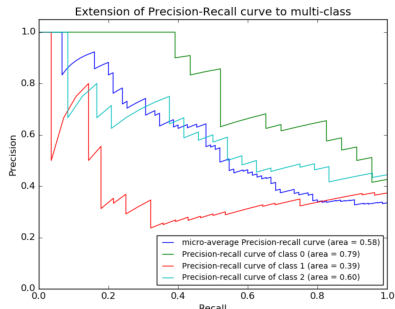
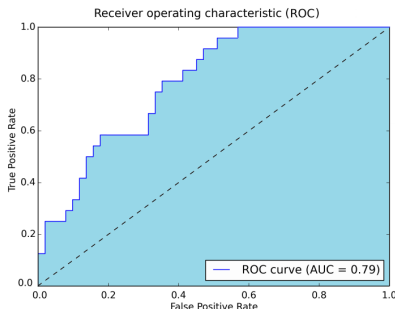
$$\text{Recall: } R = \frac{1}{|Y|} \sum_y \frac{TP_y}{TP_y + FN_y};$$

Макроусреднение чувствительно
к ошибкам на малочисленных классах



Кривые ROC и Precision-Recall

Бинарный классификатор: $a(x; w, w_0) = \text{sign}(b(x, w) - w_0)$
Каждая точка кривой соответствует значению порога w_0



AUROC — площадь под ROC-кривой

AUPRC — площадь под кривой Precision-Recall

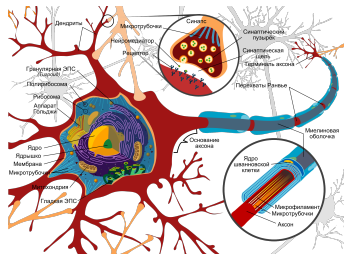
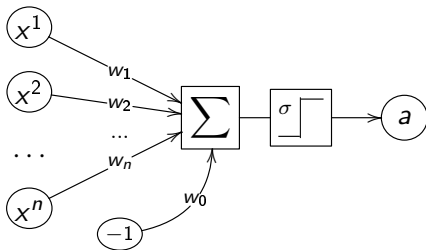
Примеры из Python scikit learn: <http://scikit-learn.org/dev>

Линейный классификатор — математическая модель нейрона

Модель МакКаллока-Питтса [1943] при $w_j \in \{-\infty, 1\}$, $\sigma = \text{sign}$;
 персептрон Розенблатта [1957] при $w_j \in \mathbb{R}$, $\sigma = \text{sign}$:

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

$\sigma(z)$ — функция активации (например, sign , th или $\frac{1}{1+e^{-z}}$),
 w_j — весовые коэффициенты синаптических связей,
 w_0 — порог активации, он же вес признака $f_0(x) \equiv -1$



Проблема переобучения в линейных моделях

$a(x, w) = \text{sign}\langle w, x \rangle$ — линейный бинарный классификатор

$a(x, w) = \langle w, x \rangle$ — линейная модель регрессии

Мультиколлинеарность — линейная зависимость признаков:

- $\exists u \in \mathbb{R}^n: \forall x \in X \quad \langle u, x \rangle = 0$
- неединственность решения: $\forall \gamma \in \mathbb{R} \quad \langle w, x \rangle = \langle w + \gamma u, x \rangle$

Это приводит к неустойчивости и переобучению:

- неустойчивость: слишком большие веса $|w_j|$ разных знаков
- переобучение: $Q(w^*, X^\ell) \ll Q(w^*, X^k)$

Как уменьшить переобучение:

- регуляризация $\|w\| \rightarrow \min$ (сокращение весов, weight decay)
- отбор признаков: $f_1, \dots, f_n \rightarrow f_{j_1}, \dots, f_{j_m}, \quad m \ll n$
- преобразование признаков: $f_1, \dots, f_n \rightarrow g_1, \dots, g_m, \quad m \ll n$

Регуляризация по L_2 -норме (сокращение весов, weight decay)

Штраф за увеличение нормы вектора весов:

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

Градиент:

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

Модификация градиентного шага в методе SG:

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

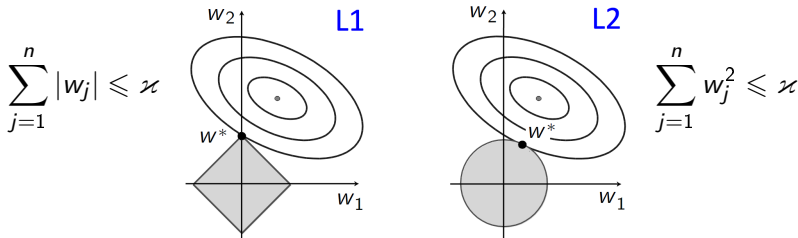
Методы подбора коэффициента регуляризации τ :

- 1 hold-out или скользящий контроль
- 2 стохастическая адаптация: чаще выбираем такие значения из сетки $\{\tau_k\}$, которые в среднем более успешны

Регуляризация по L_1 -норме для отбора признаков

LASSO — Least Absolute Shrinkage and Selection Operator

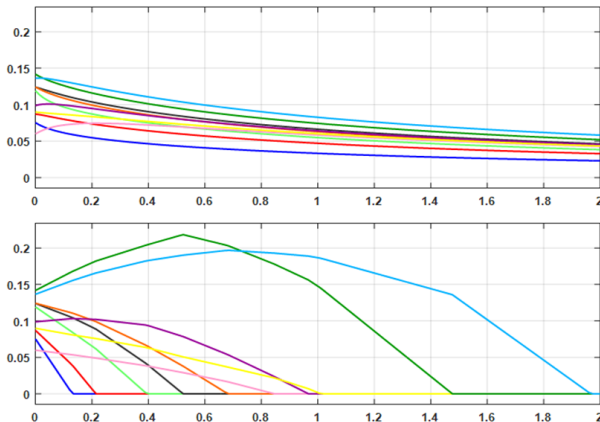
$$Q(w) + \tau \sum_{j=1}^n |w_j|^p \rightarrow \min_w \iff \begin{cases} Q(w) \rightarrow \min_w; \\ \sum_{j=1}^n |w_j|^p \leq \tau; \end{cases}$$



ElasticNet: $\tau|w_j| + \lambda|w_j|^2$ — и так тоже можно (но не сегодня)

Сравнение L_2 и L_1 регуляризации

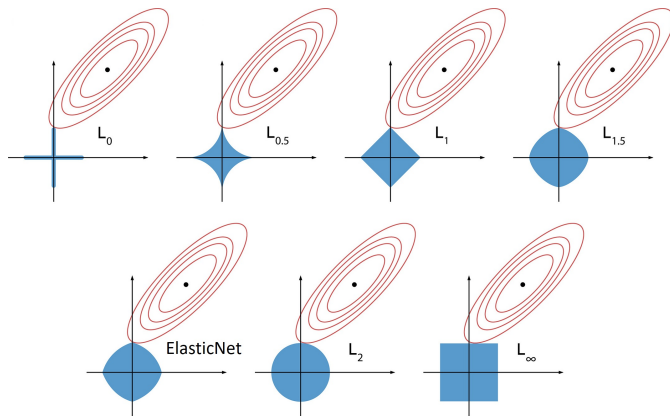
Зависимость весов w_j от параметра селективности τ



В LASSO с увеличением τ усиливается отбор признаков

Геометрическая интерпретация отбора признаков

Сравнение регуляризаторов по различным L_p -нормам:



Негладкость регуляризатора приводит к отбору признаков

Логистическая регрессия для бинарной классификации

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

Найти: параметры $w \in \mathbb{R}^n$ линейной модели классификации

$$a(x, w) = \text{sign} \sum_{j=1}^n w_j f_j(x) = \text{sign} \langle w, x \rangle$$

Критерий $\min$ аппроксимированного эмпирического риска:

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

margin

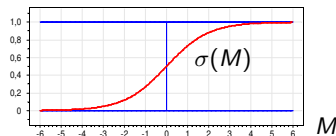
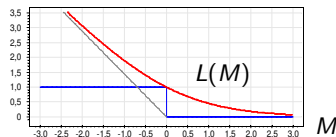
логарифмическая функция потерь:

$$[M < 0] \leq L(M) = \ln(1 + e^{-M})$$

Условные вероятности классов:

$$P(y|x, w) = \sigma(\langle w, x \rangle y)$$

$\sigma(M) = \frac{1}{1+e^{-M}}$ — сигмоидная функция



Принцип максимума правдоподобия

Пусть $X \times Y$ — в.п. с плотностью $p(x, y)$

Пусть X^ℓ — простая (i.i.d.) выборка: $(x_i, y_i)_{i=1}^\ell \sim p(x, y)$

Задача: по выборке X^ℓ оценить плотность $p(x, y)$

$p(x, y) = P(y|x, w)p(x)$ — параметрическая модель плотности:

$P(y|x, w)$ — модель условной вероятности классов, параметр w

$p(x)$ — неизвестное и непараметризуемое распределение на X

Максимум правдоподобия (Maximum Likelihood Estimate, MLE):

$$p(X^\ell, w) = \prod_{i=1}^{\ell} p(x_i, y_i) = \prod_{i=1}^{\ell} P(y_i|x_i, w) \overline{p(x_i)} \rightarrow \max_w$$

Максимум логарифма правдоподобия (log-likelihood, log-loss):

$$Q_{\text{MLE}}(w) = \sum_{i=1}^{\ell} \log P(y_i|x_i, w) \rightarrow \max_w$$

Связь правдоподобия и аппроксимации эмпирического риска

Максимизация правдоподобия в задаче классификации,
где $P(y|x, w)$ — модель условной вероятности класса y :

$$Q_{\text{MLE}}(w) = \sum_{i=1}^{\ell} \log P(y_i|x_i, w) \rightarrow \max_w$$

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

$$Q_{\text{ERM}}(w) = \sum_{i=1}^{\ell} L(y_i b(x_i, w)) \rightarrow \min_w$$

Эти два принципа эквивалентны, если положить

$$-\log P(y_i|x_i, w) = L(y_i b(x_i, w))$$

$$\boxed{\text{модель } P(y|x, w)} \iff \boxed{\text{модель } b(x, w) \text{ и } L(M)}$$

Вероятностный смысл регуляризации

$P(y|x, w)$ — вероятностная модель данных

$p(w; \gamma)$ — априорное распределение параметров модели

γ — вектор гиперпараметров

Теперь вероятностная модель описывает появление не только данных X^ℓ , но и самой задачи как модели w из W

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):

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

Примеры: априорные распределения Гаусса и Лапласа

Пусть веса w_j независимы, $E w_j = 0$, $D w_j = C$.

Распределение Гаусса и квадратичный (L_2) регуляризатор:

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

Распределение Лапласа и абсолютный (L_1) регуляризатор:

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

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

Чем меньше дисперсия, тем сильнее регуляризация

Пример. Бинаризация признаков и скоринговая карта

Задача кредитного скоринга:

- x_j — заёмщики
- $y_i = -1$ (bad), $+1$ (good)

Бинаризация признаков $f_j(x)$:

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

Линейная модель классификации:

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

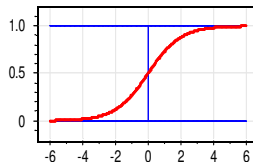
Вес признака w_{jk} равен его вкладу в общую сумму баллов (score).

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

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



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

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

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 $L(M)$*
позволяет использовать градиентную оптимизацию
- Функции $L(M)$, штрафующие за приближение к границе классов, увеличивают зазор между классами, благодаря чему повышается надёжность классификации
- *Измерения качества:* Acc, Sen, Sp, P , R , F_1 , AUC, log-loss
- *Регуляризация* снижает переобучение, возникающее в линейных моделях из-за мультиколлинеарности
- *Логистическая регрессия* — метод классификации, оценивающий условные вероятности классов $P(y|x)$