Введение в машинное обучение

Бизнес в стиле .RU

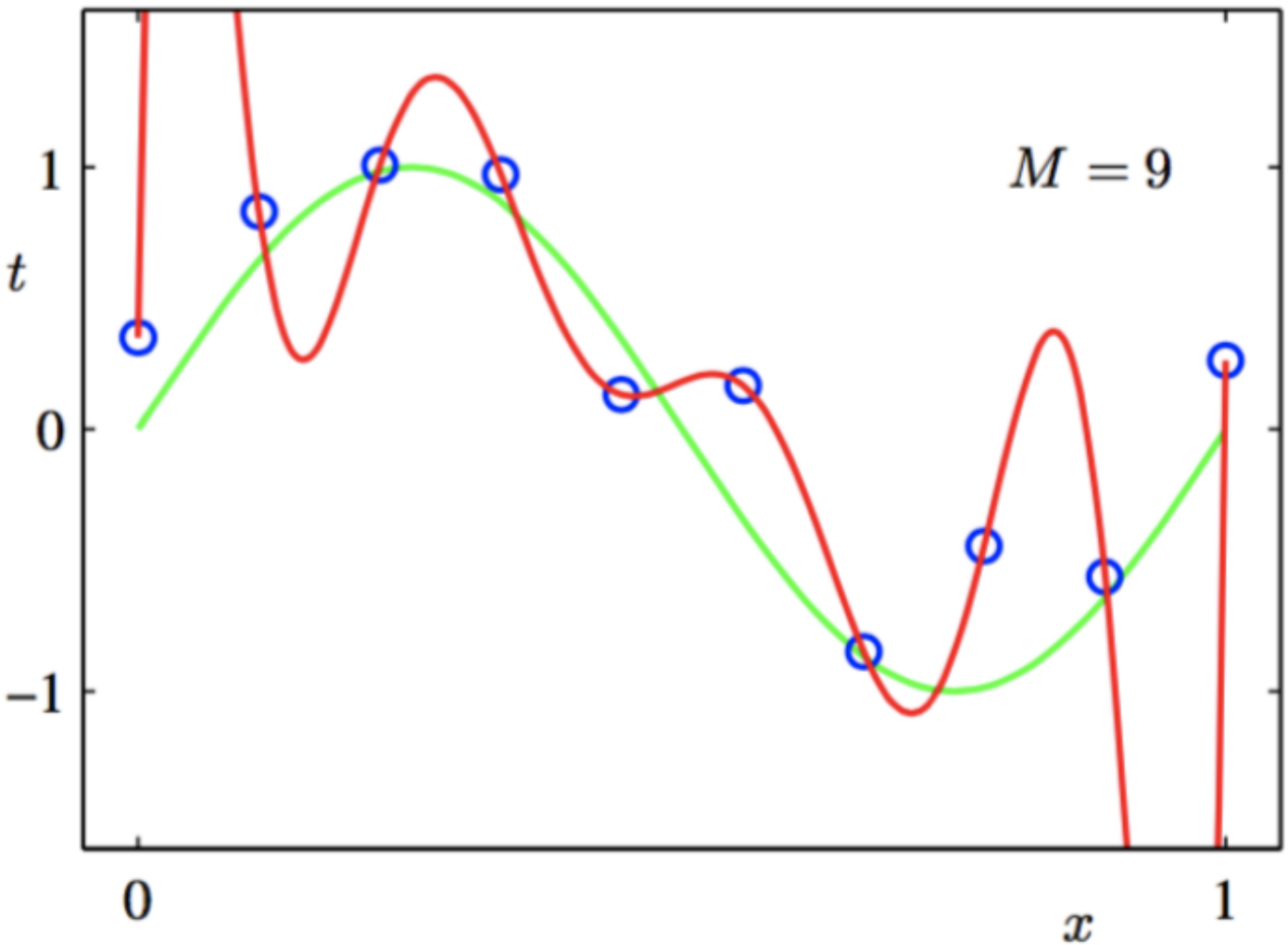
Лекция 2

Шикунов Николай
Telegram: @shik_n
nashikunov@gmail.com
НИУ ВШЭ

14 апреля 2020 г.

Переобучение

Переобучение — 9 признаков: $a(x) = w_0 + w_1 x_1 + ... + w_9 x_9$



Переобучение

- При использовании признаков высоких степеней модель получает возможность слишком хорошо подстроиться под выборку!
- Посмотрим на коэффициенты переобученной линейной модели
- $a(x) = 0.8 + 631.34x_1 + 84729.43x_2 + ... + 374219x_9$
- Эмпирически заметим, что у переобученной модели очень большие веса!

Регуляризация

Добавим к функционалу «регуляризатор», который штрафует за слишком большую норму вектора весов

- $Q_\alpha(w) = Q(w) + \alpha R(w)$
- α — гиперпараметр регуляризации (гипер?)
- Высокий параметр α — простая модель. Жёстко штрафуем и следим за переобучением
- Низкий параметр α — модель сложнее. Маленький штраф и риск переобучения сохраняется

Гиперпараметры и параметры

Параметры

Параметры настраиваются по обучающей выборке. Например: веса линейной регрессии

Гиперпараметры

Гиперпараметры контролируют сам процесс обучения. Их нельзя подобрать по обучающей выборке. Мы подбираем их на валидационной выборке или по кросс-валидации

Гиперпараметры и параметры

α — гиперпараметр регуляризации

Введение регуляризации мешает модели подгоняться под обучающие данные, и, с точки зрения среднеквадратичной ошибки, выгодно всегда брать $\alpha = 0$. Мы вводим регуляризацию, чтобы улучшить результаты модели на данных, которые она не видела. $\alpha = 0$ не делает этого. Поэтому коэффициент регуляризации (как и другие гиперпараметры) следует настраивать по отложенной выборке или с помощью кросс-валидации

L_2 - регуляризация

L_2 - регуляризация

$$R(w) = \|w\|_2 = \sum_{i=1}^{d} w_i^2$$

Обучение линейной регрессии с MSE и L_2-регуляризацией:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\left(\langle \mathbf{w}, \mathbf{x_i}\rangle - y_i\right)^2 + \alpha||w||_2 \to \min_w .$$

- Функционал гладкий и выпуклый
- $w = (X^T X + \alpha I)^{-1} X^T y$
- При добавлении диагональной матрицы к $X^T X$ данная матрица становится положительно определённой, и поэтому её можно обратить
- Решение всегда будет единственным!

L_1 - регуляризация

L_1 - регуляризация

$R(w) = \|w\|_1 = \sum_{i=1}^{d} |w_i|$

Обучение линейной регрессии с MSE и L_1-регуляризацией:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\left(\langle \mathbf{w}, \mathbf{x_i}\rangle - y_i\right)^2 + \alpha||w||_1 \to \min_w.$$

- Функционал негладкий!
- Оптимизация сложнее, нет производной в нуле
- Интересная особенность: часть весов обратятся в ноль. Происходит «отбор» признаков

Разреженные модели

Зачем нужно «отбирать» признаки?

- Модели, в которых некоторые веса равны нулю, называют разреженными, поскольку прогноз в них зависит лишь от части признаков

Зачем?

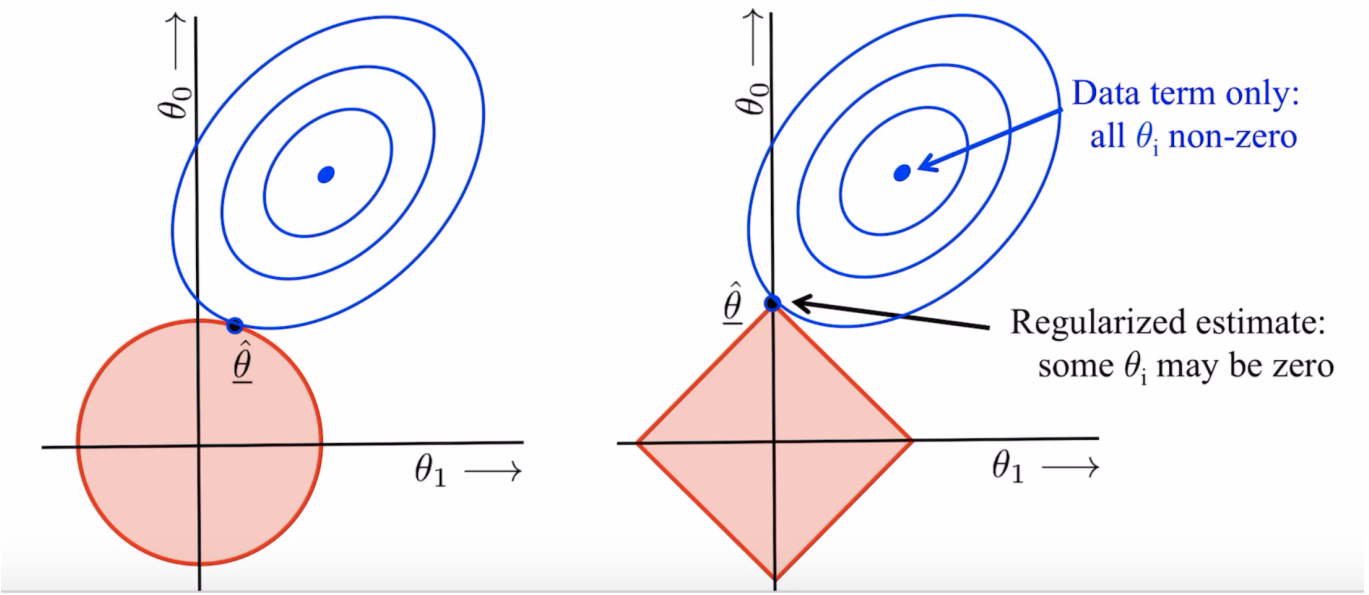
- Нет смысла использовать в модели признаки, которые не влияют на поставленную задачу/предметную область. Это лишь добавляет шум
- Признаков может быть «очень» много. Иногда возникает потребность ускорить процесс обучения/инференса модели. Поэтому требуется совершить отбор признаков

L_1-регуляризация и отбор признаков

- $Q_\alpha(w) = Q(w) + \alpha R(w)$
-

$$\begin{cases} Q(w) \to \min\limits_w \\ \|w\|_1 \leq C \end{cases}$$

- L1 tends to generate sparser solutions than a quadratic regularizer



Линейная зависимость признаков

$$\mathbf{X} = \begin{pmatrix} 1 & 10 & 6 & 3.5 \\ 2 & 20 & 0 & 1 \\ 3 & 30 & 2 & 2.5 \end{pmatrix}$$

- $x_2 = 10x_1$
- $x_4 = 0.5x_1 + 0.5x_3$

Линейная зависимость признаков

- Возникает избыточная информация
- Лишние затраты на хранение данных
- Приводит к тому, что решении бесконечное число
- Может возникать переобучение

Линейная зависимость признаков

1. Пусть в выборке есть линейно зависимые признаки
2. По определению линейной зависимости существует такой вектор v, что для любого объекта x выполнено $\langle v, x \rangle = 0$
3. Допустим, мы нашли оптимальный вектор весов w для линейной модели
4. Модифицируем наши веса (по приколу)

$$\langle w + \alpha v, x \rangle = \langle w, x \rangle + \alpha \underbrace{\langle v, x \rangle}_{=0} = \langle w, x \rangle.$$

5. Метод оптимизации может найти решение со сколько угодно большими весами! $=>$ переобучение

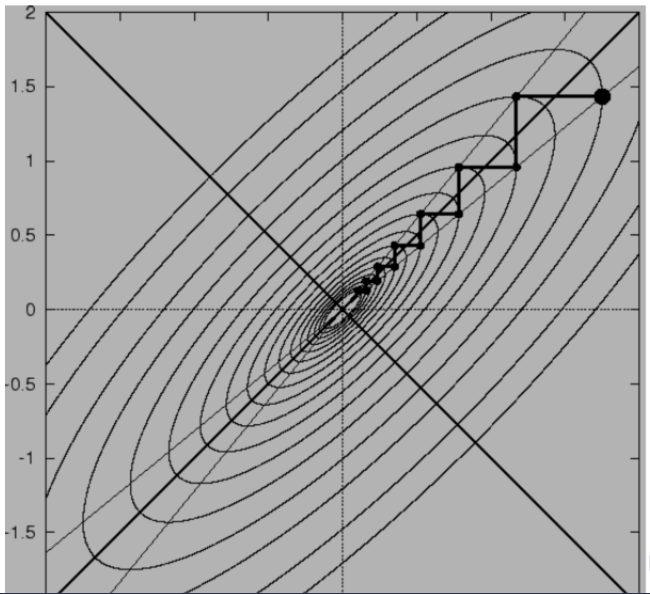
С помощью линейной регрессии можно восстанавливать нелинейные зависимости, если провести преобразование признакового пространства!

- Например, квадратичные признаки: $(x_1, \ldots, x_d, x_1^2, \ldots, x_d^2, x_1 x_2, \ldots, x_{d-1} x_d)$
- Можно работать с полиномиальными признаками больших порядков
- Аналогично $log, exp, sin...$
- Главное следить за переобучением!

Преобразование признаков

Масштабирование признаков

1. $f(x) = \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2$
2. Градиентный спуск. $x^{(0)} = (1, 1), \eta = 1$
3. Антиградиент — $(-1, -1)$. Сходимся за 1 шаг
4. Теперь «растянем» функцию вдоль одной из осей: $f(x) = 50x_1^2 + \frac{1}{2}x_2^2$
5. Градиентный спуск. $x^{(0)} = (1, 1), \eta = 1$
6. Антиградиент — $(-100, -1)$. За 1 шаг уже не сходимся, да и сойдёмся ли вообще...



Аналогичная проблема возникает с функционалом ошибки в линейной регрессии, если один из признаков существенно отличается по масштабу от остальных!

- x_1 — год рождения пользователя
- x_2 — траты пользователя за 2 месяца
- x_3 — количество детей у пользователя

Стандартизация признаков

$$x_{ij} := \frac{x_{ij} - \mu_j}{\sigma_j},$$

где μ — среднее значение, σ — стандартное отклонение

Масштабирование признаков на отрезок $[0, 1]$

$$x_{ij} := \frac{x_{ij} - \min_i x_{ij}}{\max_i x_{ij} - \min_i x_{ij}}.$$

Линейная классификация

- $Y = \{-1, +1\}$
- класс «$+1$» положительный класс
- класс «-1» отрицательный класс

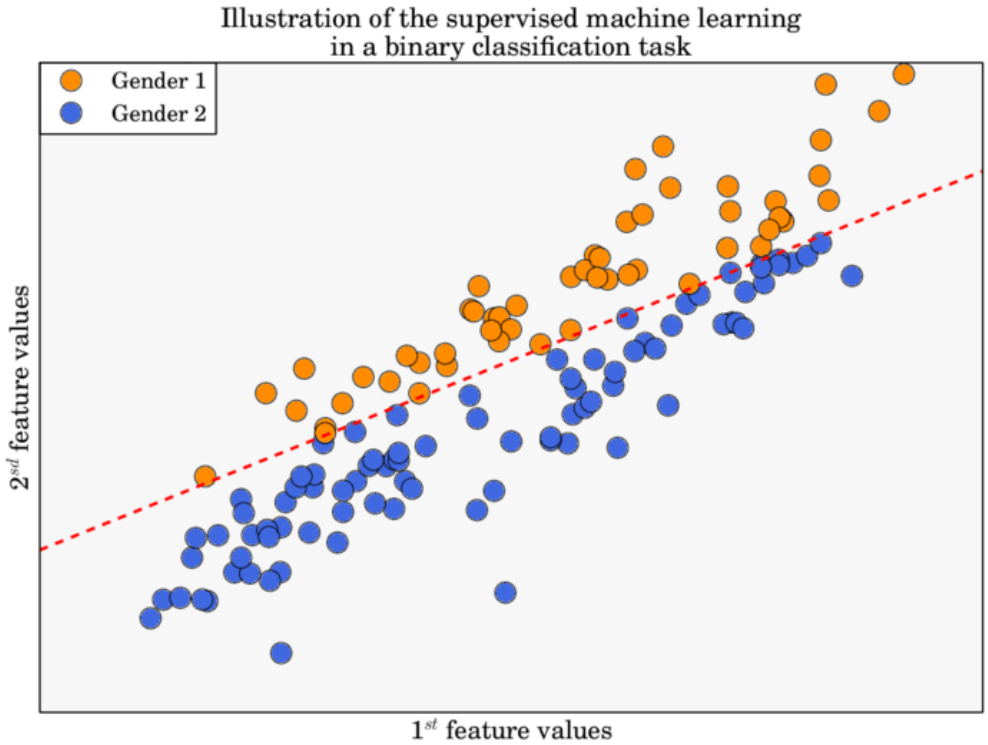
Линейная модель линейного классификатора

$$a(x) = (\langle w, x \rangle + w_0) = sign\left(\sum_{j=1}^{d} w_j x_j + w_0\right) = sign\langle w, x \rangle.$$

Линейная классификация

- Геометрически линейный классификатор соответствует гиперплоскости с вектором нормали w
- $\langle w, x \rangle$ пропорционально расстоянию от гиперплоскости до точки x
- Знак показывает с какой стороны от гиперплоскости находится данная точка

Линейная классификация

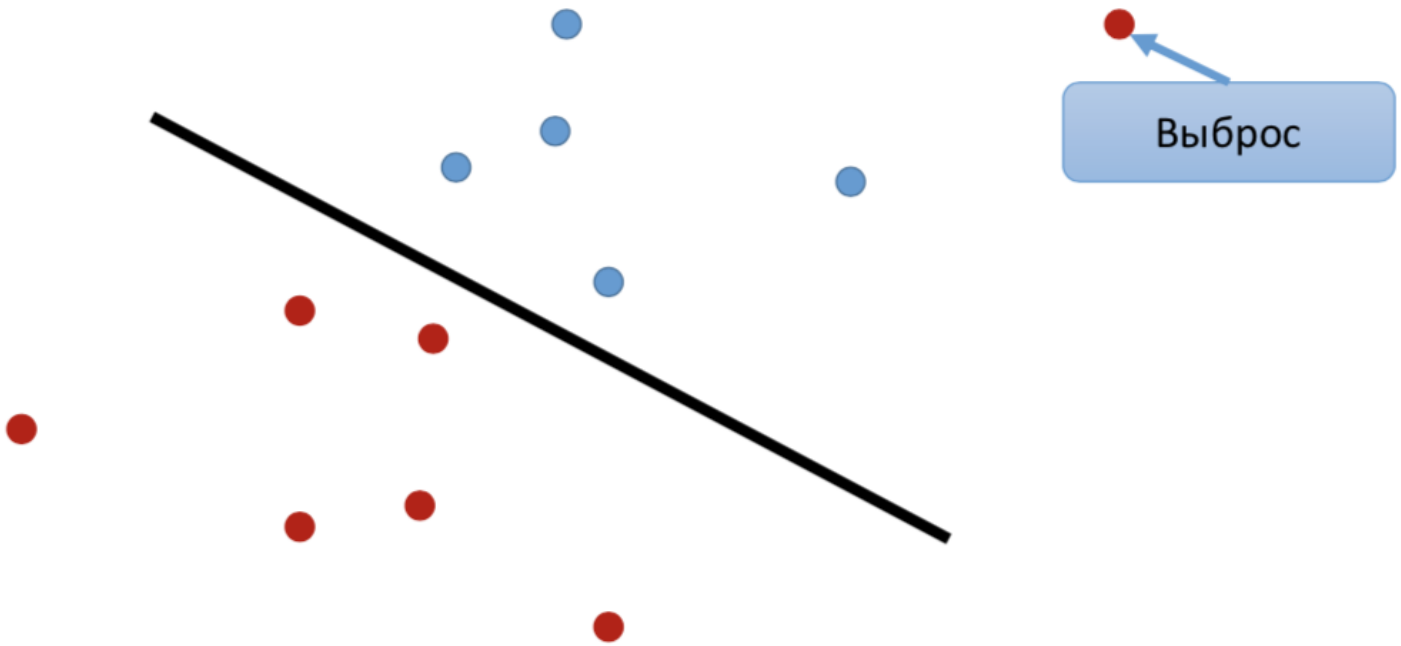


- $\langle w, x \rangle < 0$ — объект «слева» от прямой
- $\langle w, x \rangle > 0$ — объект «справа» от прямой

Линейная классификация

- Расстояние от точки до гиперплоскости: $\frac{|\langle w,x\rangle|}{||w||}$
- Чем больше $\langle w, x\rangle$, тем дальше объект от разделяющеи гиперплоскости

Линейная классификация



Линейная классификация

Отступ

- $M_i = y_i \langle w, x_i \rangle$
- $M_i > 0$ — классификатор даёт верный ответ
- $M_i < 0$ — классификатор ошибается
- Чем больше $|M_i|$, тем больше уверенность классификатора в ответе

Доля неправильных ответов

- $Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) \neq y_i] = \frac{1}{\ell} \sum_{i=1}^{\ell} [\langle w, x_i \rangle \neq y_i] \to \min_w$
- Запишем через отступ
-

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [\underbrace{y_i \langle w, x_i \rangle}_{M_i} < 0] \to \min_w$$

Пример

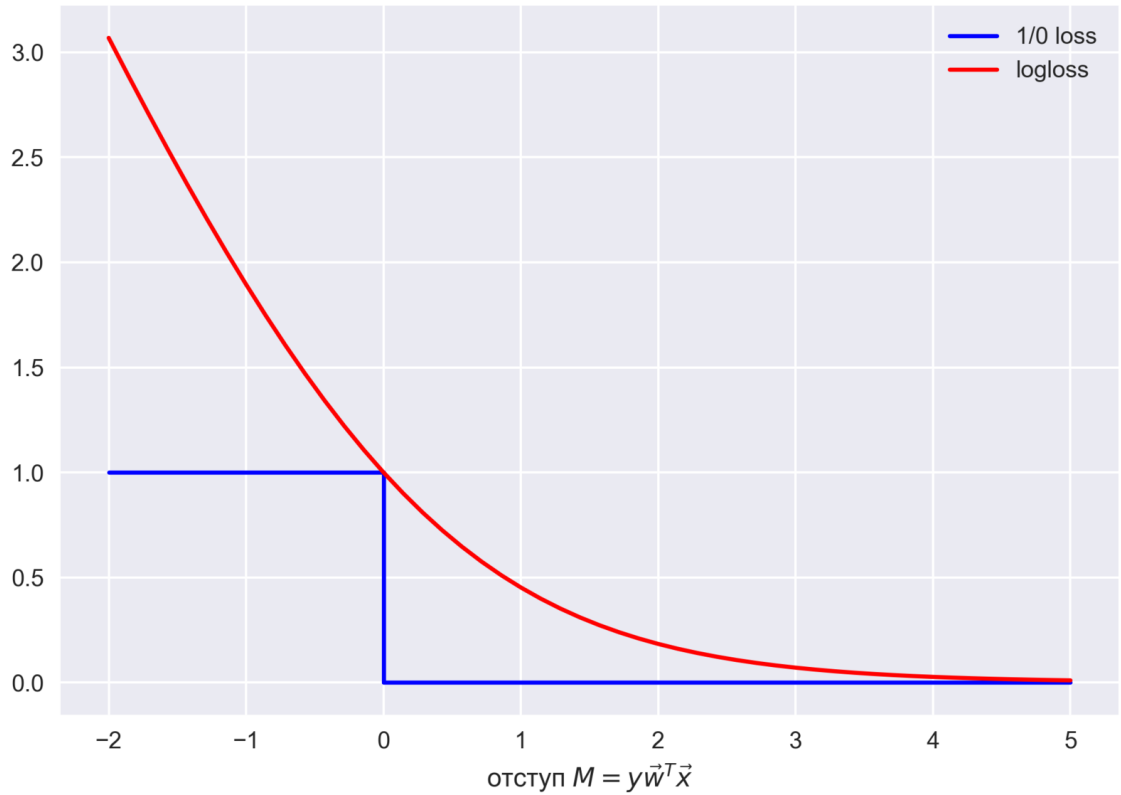
- $\mathbf{y} = (1, 1, 1, 0, 0, 1)$
- $a(x) = (1, 1, 1, 1, 1, 1)$
- Доля неправильных ответов $= \frac{2}{6} = 0.33$

Доля неправильных ответов

- $Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [\underbrace{y_i \langle w, x_i \rangle}_{M_i} < 0] \to \min_w$
- Разрывная функция
- Как это оптимизировать?

Линейная классификация

Пороговая функция потерь



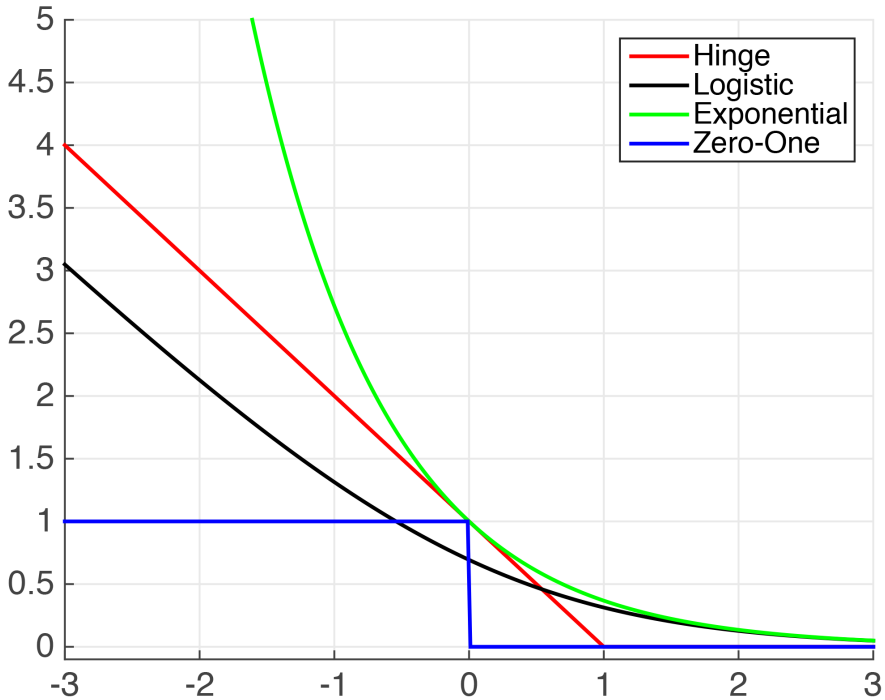
- Возьмем любую гладкую оценку пороговои функции
- $L(M) = [M < 0] \leq \tilde{L}(M)$, где $M = y\langle w, x\rangle$
- После этого можно получить верхнюю оценку

$$Q(a, X) \leq \frac{1}{\ell} \sum_{i=1}^{\ell} \tilde{L}(y_i \langle w, x_i \rangle) \to \min_w$$

Линейная классификация

Функционал качества

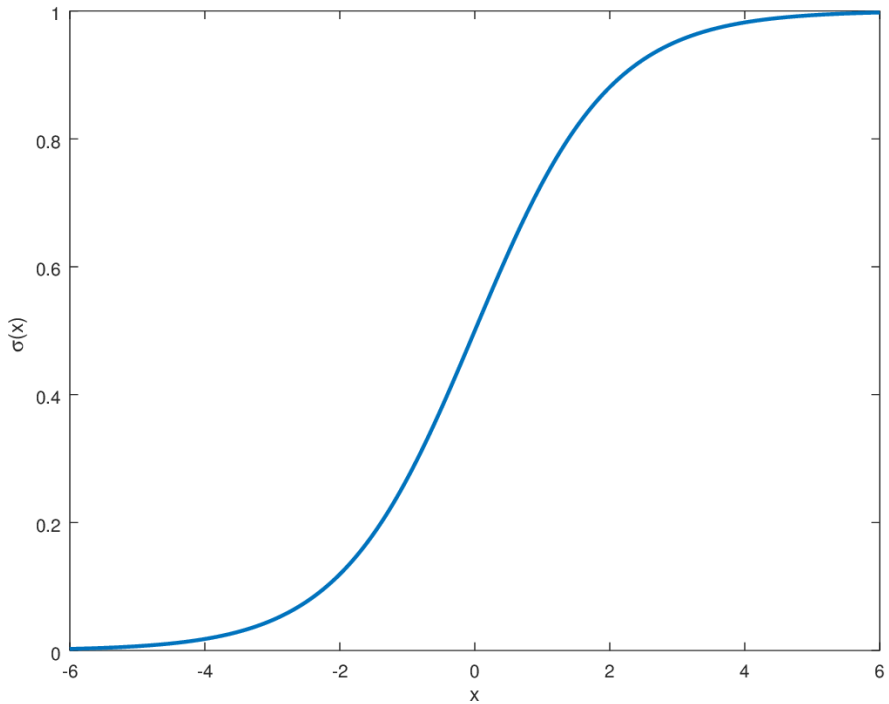
1. $\tilde{L}(M) = \log\left(1 + e^{-M}\right)$ — Logistic Loss
2. $\tilde{L}(M) = (1 - M)_+ = \max(0, 1 - M)$ — Hinge Loss (для SVM)
3. $\tilde{L}(M) = e^{-M}$ — Exponential Loss
4. $\tilde{L}(M) = 2/(1 + e^{M})$ — Sigmoid Loss
5. ...



- $Q(a, X) = \sum_{i=1}^{\ell} \log\left(1 + \exp(-y_i \langle w, x_i \rangle)\right)$
- В чём её особенность?
- Логистическая регрессия позволяет оценивать вероятности принадлежности к классам!

- Сигмоидная функция: $\sigma(z) = \frac{1}{1+\exp(-z)}$
-

$$p(y = 1|x) = \frac{1}{1 + \exp(-\langle w, x \rangle)}.$$



Метрики качества классификации

Accuracy (Доля правильных ответов)

$$\text{accuracy}(a, x) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) = y_i].$$

- Самая очевидная и простая метрика
- Не переводить, как «точность»!

Метрики качества классификации

Accuracy (Доля правильных ответов)

- Допустим в выборке 1000 объектов
- 950 объектов положительного класса, 50 объектов отрицательного класса
- Пусть $a(x) = 1$ (по приколу)
- Получается, что $Accuracy = 0.95$!
- Если наша выборка несбалансированная, то Accuracy использовать нельзя
- Если получили большои accuracy — посмотрите на баланс классов

Метрики качества классификации

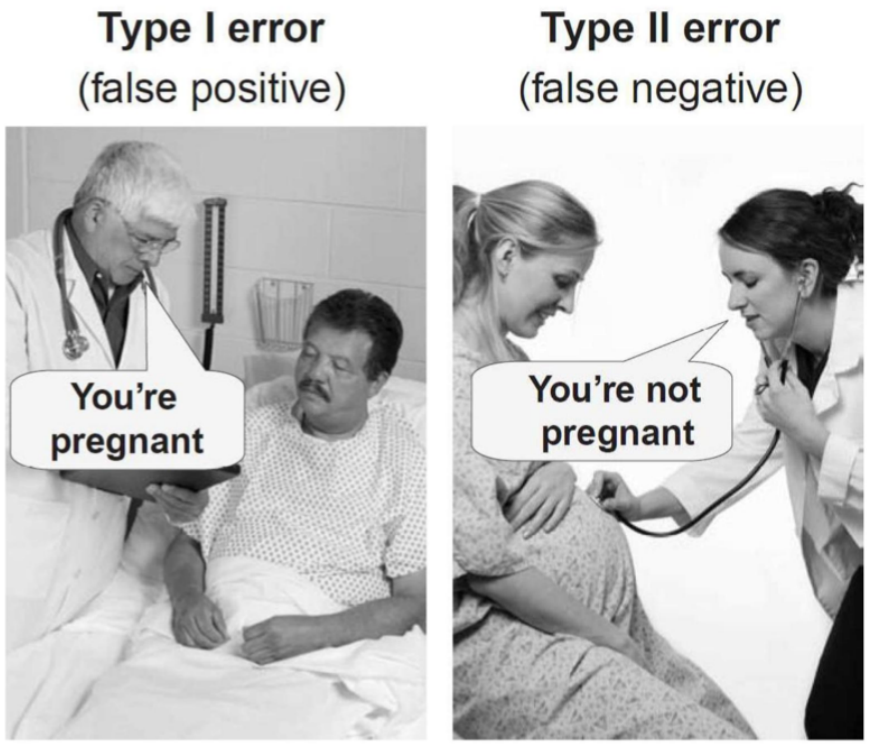
Матрица ошибок (confusion matrix)

	$y = 1$	$y = -1$
$a(x) = 1$	True Positive (TP)	False Positive (FP)
$a(x) = -1$	False Negative (FN)	True Negative (TN)

- False Positive - Ошибка первого рода (ложная тревога)
- False Negative - Ошибка второго рода (пропуск цели)

Метрики качества классификации

Матрица ошибок (confusion matrix)



Метрики качества классификации

Точность (precision) и Полнота (recall)

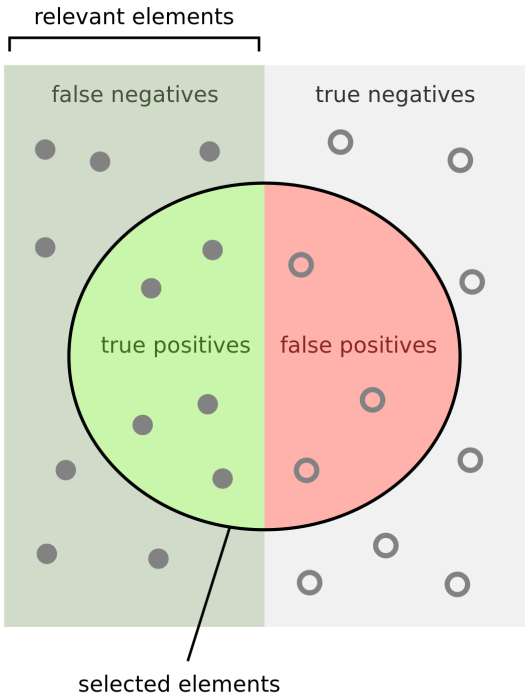
$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}};$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

- Точность показывает, какая доля объектов, выделенных классификатором как положительные, действительно является положительными
- Полнота показывает, какая часть положительных объектов была выделена классификатором

Метрики качества классификации

Точность (precision) и Полнота (recall)



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{selected elements}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

Точность и полнота устойчивы к несбалансированным выборкам!

Предсказание оттока сотрудников компании

Модель должна предсказывать увольнение сотрудника за 1 месяц до того, как он напишет заявление. Менеджер ежемесячно получает список людей, которые скорее всего уволятся. Его задача поговорить с каждым сотрудником из списка и попробовать удержать.

Метрики качества классификации

Точность (precision) и Полнота (recall)

	$y = 1$ Сотрудник уволился через месяц	$y = -1$ Сотрудник не уволился через месяц
$a(x) = 1$ Сотрудник уволится	True Positive (TP) 9	False Positive (FP) 50
$a(x) = -1$ Сотрудник не уволится	False Negative (FN) 1	True Negative (TN) 40

- $precision = \frac{9}{9+50} = 0.152$
- $recall = \frac{9}{9+1} = 0.9$
- Модель выявила 0.9 случаев потенциальных увольнений!
- Но менеджер 59 раз ходил раговаривать с сотрудниками, пытался их удержать. А это время и деньги
- Другой пример: выявлен злокачественной опухоли у больных клиники
- Мы должны балансировать между точностью и полнотой в зависимости от специфики задачи!

Метрики качества классификации

F-мера

$$F = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}.$$

- Гармоническое среднее точности и полноты
- F-мера близка к нулю, если хотя бы один из аргументов близок к нулю
- F-мера из прошлой задачи 0.26

Метрики качества классификации

F-мера

$$F_\beta = (1 + \beta^2)\frac{2 * \text{precision} * \text{recall}}{\beta^2 \text{precision} + \text{recall}}.$$

- β принимает значения в диапазое $0 < \beta < 1$, если вы хотите отдать приоритет точности, а при $\beta > 1$ приоритет отдается полноте

Оценки принадлежности классу

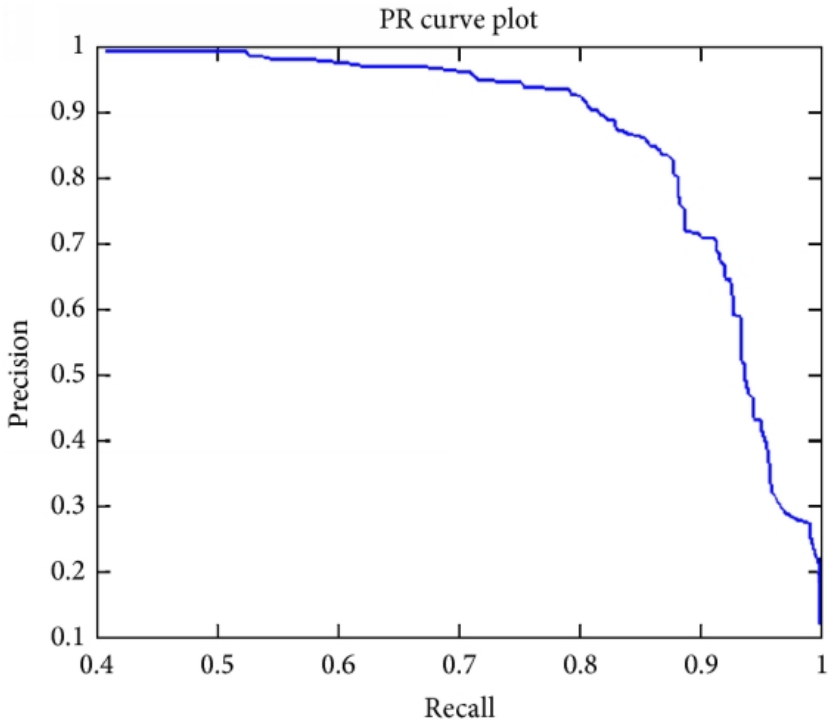
- $a(x) = [b(x) > t]$
- $b(x)$ - оценка принадлежности классу +1
- Линейный классификатор: $b(x) = \langle w, x_i \rangle$
- Как подобрать порог?

Оценки принадлежности классу

- Допустим задача кредитного скоринга - бинарная классификация
- В базовом случае порог 0.5
- Допустим $b(x_i) = 0.45$, а наш порог 0.4 $=> a(x_i) = 1$

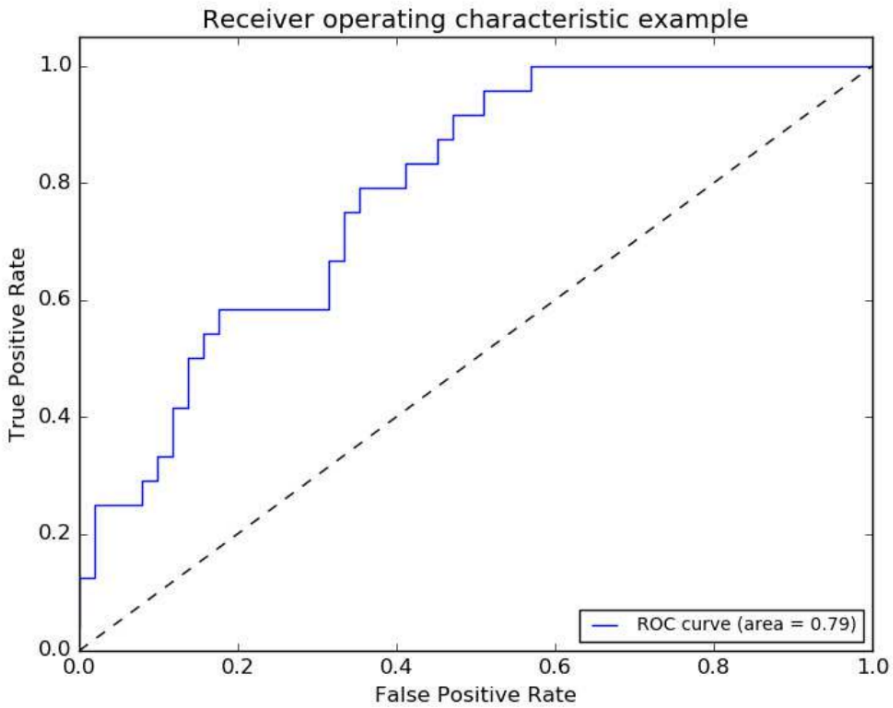
Метрики качества классификации

Precision-Recall Curve



1. Перебором смотрим все возможные пороги и считаем точность и полноту
2. Метрика AUC-PRC - площадь под кривой PR

Метрики качества классификации
AUC-ROC



1. Перебором смотрим все возможные пороги и считаем TPR и FPR
2. Метрика AUC-ROC - площадь под кривой ROC

Метрики качества классификации
AUC-ROC

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}};$$
$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

- False Positive Rate, FPR - доля неверно принятых объектов
- True Positive Rate, TPR - доля верно принятых объектов
- Идеальныи алгоритм: $AUCROC = 1$
- Если модель случайно классифицирует: $AUCROC = 0.5$

Метрики качества классификации
AUC-ROC

- Метрики нашей модели: $F_1 = 0.34, AUC - ROC = 0.89$
- Плохой классификатор? Нет!
- Это говорит о том, что нужно подобрать порог!
- Пример: Если вероятность увольнения сотрудника больше 0.35, тогда модель выдаст ответ 1

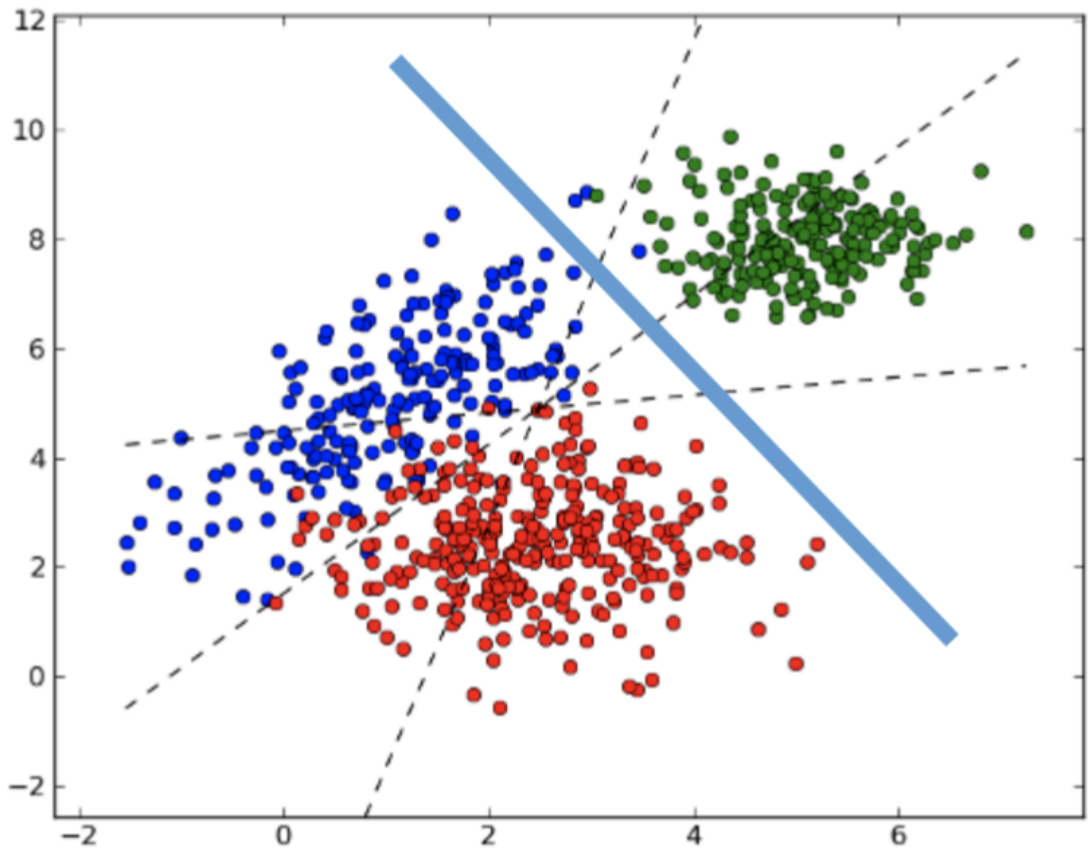
Многоклассовая классификация

One-vs-All

- Способ сведения многоклассовых задач к набору бинарных классификации
- Обучаем классификатор для каждого класса
- Задача: отделение класса от всех остальных
- k классов $=>$ k задач классификации
- $a(x) = argmax_{k \in \{1,\dots,K\}} b_k(x)$ - самый уверенный класс

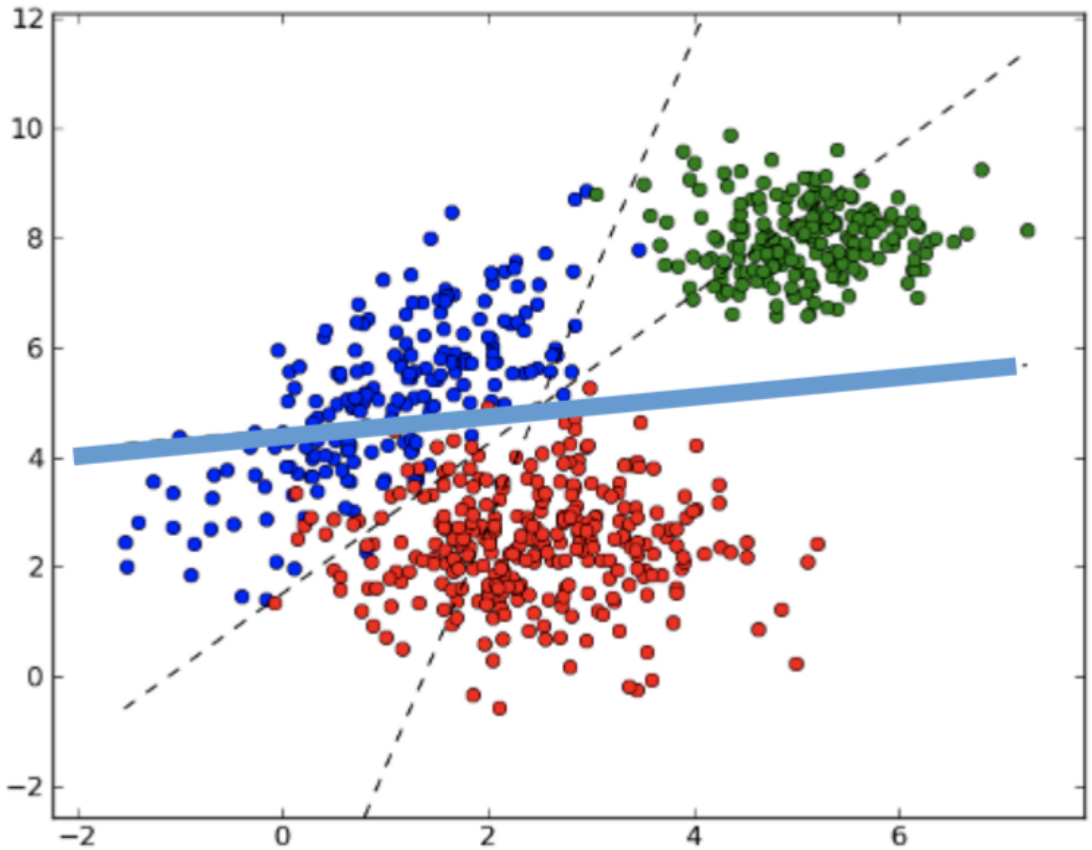
Многоклассовая классификация

One-vs-All



Многоклассовая классификация

One-vs-All



Проблема данного подхода заключается в том, что каждый из классификаторов $b_1(x), \ldots, b_K(x)$ обучается на своей выборке, и выходы этих классификаторов могут иметь разные масштабы, из-за чего сравнивать их будет неправильно

Многоклассовая классификация

All-vs-All

- Обучим C_K^2 классификаторов $a_{ij}(x)$, $i, j = 1, \ldots, K$, $i \neq j$ (все пары классов)
- Каждый раз будем обучаться на выборке X_ij, в которую входят только объекты классов ij
- Соответственно, классификатор $a_{ij}(x)$ будет выдавать для любого объекта либо класс i, либо класс j
- В качестве ответа выберем тот класс, за который наберется больше всего голосов
- $a(x) = _{k \in \{1,\ldots,K\}} \sum_{i=1}^{K} \sum_{j \neq i} [a_{ij}(x) = k]$
- Классификаторы обучаются на маленьких подвыборках

Многоклассовая классификация

Метрики

- Рассмотрим K задач отделения одного из классов от остальных (one-vs-all)

Микро-усреднение

- Найдем матрицу ошибок для каждой задачи (TP, FP, TN, FN)
- Усредним их по всем задачам
- Вычислим итоговые метрики
- Из-за усреднения вклад каждого класса зависит от его размера! Плохо!

Макро-усреднение

- Вычислим метрики по каждой из задач
- Усредним их по всем классам
- Все классы вносят равный вклад

Многоклассовая классификация

	TP	FP	FN	TN
$y = 1$	900	120	100	930
$y = 2$	850	70	150	980
$y = 3$	10	100	40	1900

Чему равна точность (precision)?

Микро-усреднение:

$\overline{TP}$	$\overline{FP}$	$\overline{FN}$	$\overline{TN}$
586.7	96.7	96.7	1270

Точность: 86%

Макро-усреднение:

Класс 1	Класс 2	Класс 3
88%	92%	9%

Точность: 63%