Машинное обучение и разработка данных
Методы классификации

Игнатов Дмитрий Игоревич

Факультет компьютерных наук
Департамент анализа данных и искусственного интеллекта

2016

Обзор методов

На этой лекции:

1. Задача классификации
2. Метод 1-Rule
3. Метрические методы классификации
   - Метод k ближайших соседей
4. Байесовские методы классификации
   - Наивный байесовский классификатор (Naïve Bayes Classifier)
5. Логистическая регрессия
6. Меры качества классификации
   - Точность, полнота, F-мера
   - ROC-кривая и AUC

В ближайшем будущем, возможно:

1. Метод опорных векторов (SVM)
2. Нейронные сети (Artificial Neural Networks)

План лекции

Задача обучения по прецедентам (ОП)

Постановка задачи

Дано:
- Множество **объектов** X
- Множество **меток** классов (**ответов**) Y
- **Целевая функция** $y : X \to Y$, заданная лишь в конечном множестве точек (**обучающей выборке**)

Задача: Построить (обучить) алгоритм $a : X \to Y$, экстраполирующий целевую функцию $y(x)$ (т.е. восстанавливающий функцию не только на обучающем множестве, но и на всем).

Классификация

Если $Y = \{1, 2, \ldots, I\}$, то задачу ОП называют задачей классификации на I непересекающихся классов.

Примеры задач

- Классификация документов
- Анализ тональности текста (Opinion Mining, Sentiment Analysis)
- Задача выявления спама (Spam/Ham)
- Медицинская диагностика: определение заболевания, выбор лечения, длительность и исход заболевания, и т.д.
- Синтез лекарств и предсказание токсичности соединений
- Предсказание ухода (оттока) клиентов
- Поиск месторождений в геологии
- ...

Функция потерь

Функция потерь (loss function)

Функция $L(a(x), y(x))$ характеризует величину ошибки алгоритма a на объекте x.
В задаче классификации обычно

$$L(a(x), y(x)) = \begin{cases} 1, & a(x) \neq y(x); \\ 0, & a(x) = y(x). \end{cases}$$

Эмпирический риск

Задача классификации (обучения по прецедентам) как задача оптимизации имеет вид

$$\frac{1}{n} \sum_{t=1}^{n} L(a(x_t), y(x_t)) \underset{a}{\rightarrow} min$$

В случае известной плотности распределения $p(x, y)$ возможна формулировка

$$\iint_{X \times Y} L(a(x), y(x)) p(x, y) dx dy \underset{a}{\rightarrow} min$$

Обучение, подбор параметров, оценка качества

Скользящий контроль (cross-validation)

Все множество прецедентов обычно разделяют на 2 (или 3) непересекающихся подмножества:

- X^{train} Множество для обучения (Training Set)
  - Непосредственное обучение алгоритма
  - Оценка некоторых параметров
- X^{test} Контрольное множество (Test Set)
  - Оценка качества
- X^{valid} Множество проверки (Validation Set)*
  - Оценка гиперпараметров алгоритма

Замечание: При разбиении выборки на подмножества важно сохранить «пропорции классов».

План лекции

Метод 1-Rule

[Witten et al., Data Mining, 2011]

Алгоритм

For each attribute,
  For each value of that attribute, make a rule as follows:
    count how often each class appears
    find the most frequent class
    make the rule assign that class to this attribute value.
  Calculate the error rate of the rules.
Choose the rules with the smallest error rate.

План лекции

1. Задача классификации
2. Метод 1-Rule
3. Метрические методы классификации
   - Метод k ближайших соседей
4. Байесовские методы классификации
   - Наивный байесовский классификатор (Naïve Bayes Classifier)
5. Логистическая регрессия
6. Меры качества классификации
   - Точность, полнота, F-мера
   - ROC-кривая и AUC

Метод k ближайших соседей

Пусть объекты $x_i \in X$ задаются некоторым признаковым описанием с помощью матрицы $F_{n\times m} = [f_j(x_i)] = \begin{pmatrix} f_1(x_1) & \cdot & f_m(x_1) \\ \cdot & \cdot & \cdot \\ f_1(x_n) & \cdot & f_m(x_n) \end{pmatrix}$

k-Nearest Neighbors

Пусть на множестве X задана функция расстояния $d : X \times X \to [0, \infty)$.
На обучающем множестве известна зависимость $y : X^{train} \to Y$
Идея: Схожие объекты принадлежат одному и тому же классу.

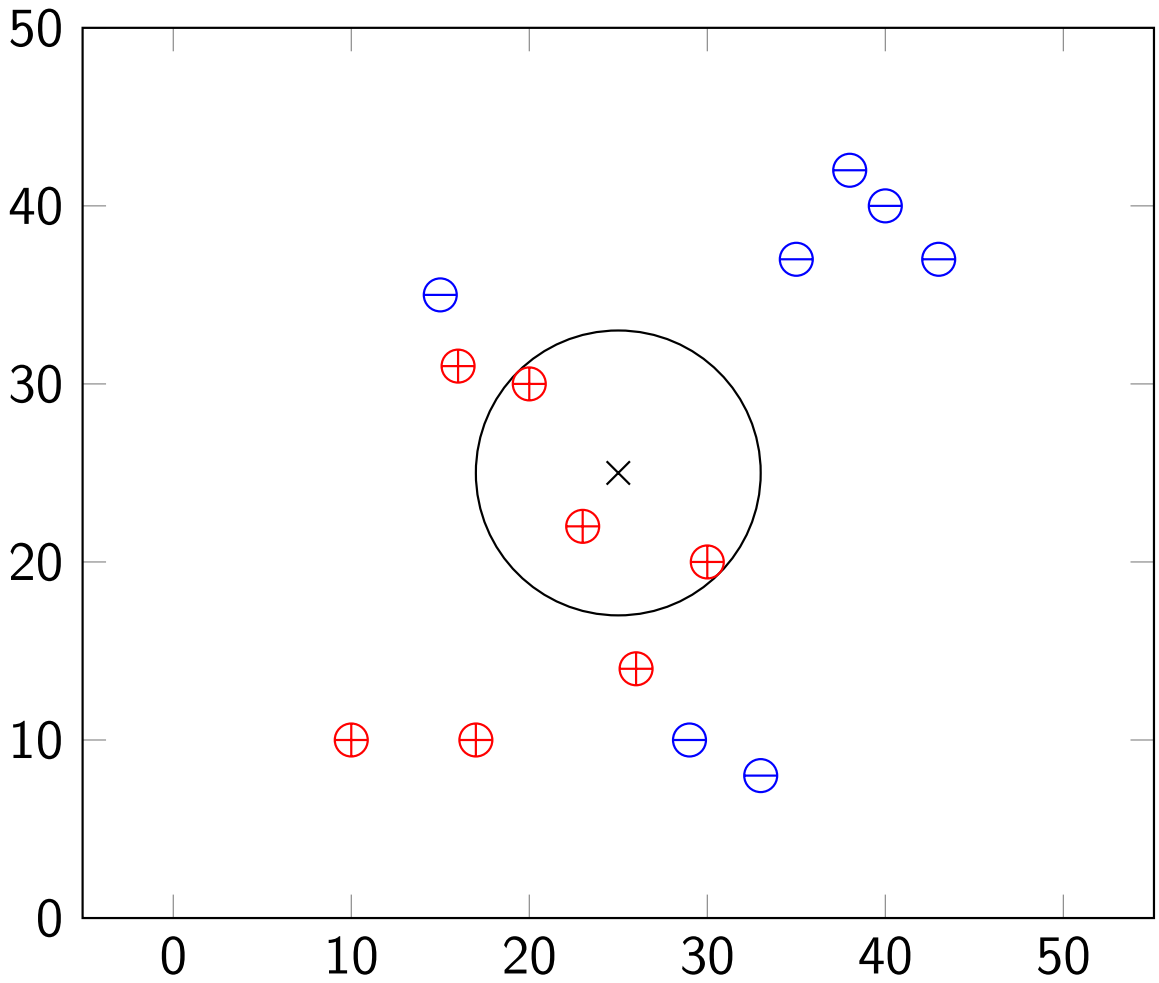
Для произвольного объекта $v \in X$ проранжируем объекты из X^{train} в порядке возрастания расстояния до v:

d	$d(v, x^{(1)})$	$\leq$	$d(v, x^{(2)})$	$\leq$	$\ldots$	$\leq$	$d(v, x^{(k)})$	$\leq$	$\ldots$
$y(x)$	1		2		$\ldots$		2		$\ldots$

Объект v будет отнесен к тому классу, элементов которого окажется больше среди k ближайших соседей

Пример

Метод ближайшего соседа $k = 3$



x – объект неопределенного класса

Метод k ближайших соседей

Вариации алгоритма

- Алгоритм ближайшего соседа ($k = 1$)
- Алгоритм k взвешенных ближайших соседей
  - Например линейно убывающие веса $w_i = \frac{k+1-i}{k}$

За и против

+ Простота реализации
– Требует хранения всего множества наблюдений
– Большая вычислительная сложность
– Неустойчив к погрешностям и выбросам в данных $\rightarrow$ эталонные объекты

План лекции

1. Задача классификации
2. Метод 1-Rule
3. Метрические методы классификации
   - Метод k ближайших соседей
4. Байесовские методы классификации
   - Наивный байесовский классификатор (Naïve Bayes Classifier)
5. Логистическая регрессия
6. Меры качества классификации
   - Точность, полнота, F-мера
   - ROC-кривая и AUC

Формула Байеса, формула полной вероятности

Формула Байеса

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\,\mathrm{P}(A)}{\mathrm{P}(B)}$$

Формула полной вероятности

$$\mathrm{P}(B) = \sum_i \mathrm{P}(A_i)\,\mathrm{P}(B|A_i)$$

Задача 1

По статистике, в половине случаев менингит провоцирует боли в шее. При этом доля пациентов с менингитом составляет $1/50000$, а на скованность мышц шеи жалуется $1/20$ всех пациентов. Является ли боль в шее значимым симптомом менингита?

Формула Байеса, формула полной вероятности

Задача 2 (Парадокс Монти-Холла)

Представьте, что вы участвуете в шоу. Перед вами 3 двери. За одной из них кроется навороченный мотоцикл, а за двумя остальными — козы. Необходимо выбрать одну из дверей. После недолгих размышлений, вы выбрали одну из дверей. Ведущий шоу, в свою очередь, открывает одну из оставшихся двух дверей, и показывает, что за ней скрыта коза, и предоставляет вам возможность ~~на ней уехать~~ изменить свой выбор. Что выгоднее сделать?

Байесовские методы классификации

Наивный байесовский классификатор (Naïve Bayes)

Пусть множество объектов X с метками классов $Y = \{y_1, y_2, \ldots, y_k\}$ описывается признаками $\{a_1, a_2, \ldots, a_m\}$.
Предположение: признаки $\{a_1, a_2, \ldots, a_m\}$ независимы.
Необходимо классифицировать новый объект x^* со значениями признаков $\{a_1^*, a_2^*, \ldots, a_m^*\}$:

$$y^* = \arg\max_{c \in Y} \mathrm{P}(y(x_i) = c | a_1^*, a_2^*, \ldots, a_m^*)$$

Величина $\mathrm{P}(y(x_i) = c | a_1^*, a_2^*, \ldots, a_m^*)$ вычисляется на множестве объектов X по формуле Байеса с учётом независимости признаков:

$$\mathrm{P}(y(x_i) = c | a_1^*, a_2^*, \ldots, a_m^*) = \frac{\mathrm{P}(a_1^*, a_2^*, \ldots, a_m^* | y(x_i) = c)\,\mathrm{P}(y(x_i) = c)}{\mathrm{P}(a_1^*, a_2^*, \ldots, a_m^*)}$$

$$\mathrm{P}(a_1^*, a_2^*, \ldots, a_m^* | y(x_i) = c) = \prod_j \mathrm{P}(a_j^* | y(x_i) = c)$$

Задача

Дана таблица

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Dl	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
Dl0	Rain	Mild	Normal	Weak	Yes
Dl1	Sunny	Mild	Normal	Strong	Yes
Dl2	Overcast	Mild	High	Strong	No

Стоит ли играть в теннис, если сегодняшний день выдался таким: $\{sunny, cool, high, strong\}$?

Сглаживание

Что если в $\prod_j \mathrm{P}(a_j^*|y(x_i)=c)$ найдется такое j', для которого $\mathrm{P}(a_{j'}^*|y(x_i)=c)=0$?

Аддитивное сглаживание (Additive Smoothing)

$$\mathrm{P}_{add}(a_j|y(x_i)) = \frac{N_{a_j}^{y(x_i)} + \lambda}{N^{y(x_i)} + \lambda V},$$

где $N_{a_j}^{y(x_i)}$ – количество объектов со значением признака a_j в классе $y(x_i)$, V — количество различных значений признака a_j, а $\lambda > 0$.

- Возможна переоценка вероятностей
- λ необходимо оптимизировать, но часто берут $\lambda = 1$

План лекции

1. Задача классификации
2. Метод 1-Rule
3. Метрические методы классификации
   - Метод k ближайших соседей
4. Байесовские методы классификации
   - Наивный байесовский классификатор (Naïve Bayes Classifier)
5. Логистическая регрессия
6. Меры качества классификации
   - Точность, полнота, F-мера
   - ROC-кривая и AUC

Логистическая регрессия

Пусть множество объектов X разбито на два класса $Y = \{0, 1\}$. Например, есть данные о размерах раковой опухоли и её доброкачественности $(y = 1)$ и недоброкачественности $(y = 0)$.

Формулировка гипотезы

Необходимо по данному признаковому описанию x найти вероятность доброкачественности опухоли, то есть $p_i = \mathrm{P}(y(x_i) = 1|x_i)$.
Определяют логистическую функцию вида $logit(p_i) = \log(\frac{p_i}{1-p_i})$.

Логистическая регрессия

Пусть множество объектов X разбито на два класса $Y = \{0, 1\}$. Например, есть данные о размерах раковой опухоли и её доброкачественности ($y = 1$) и недоброкачественности ($y = 0$).

Формулировка гипотезы

Необходимо по данному признаковому описанию x найти вероятность доброкачественности опухоли, то есть $p_i = \mathrm{P}(y(x_i) = 1|x_i)$.
Определяют логистическую функцию вида $logit(p_i) = \log(\frac{p_i}{1-p_i})$.

$$\log(\frac{p_i}{1-p_i}) = \beta^T x_i \Leftrightarrow \frac{p_i}{1-p_i} = \exp(\beta^T x_i) \Leftrightarrow p_i = \frac{1}{1+\exp(-\beta^T x_i)},$$

$$h(\beta, x_i) = \frac{1}{1+\exp(-\beta^T x_i)}$$

$h(\beta, x_i) = 0.7 \Leftrightarrow$ метка i-го объекта есть $y(x_i) = 1$ с вероятностью 0.7.

Сигмоид

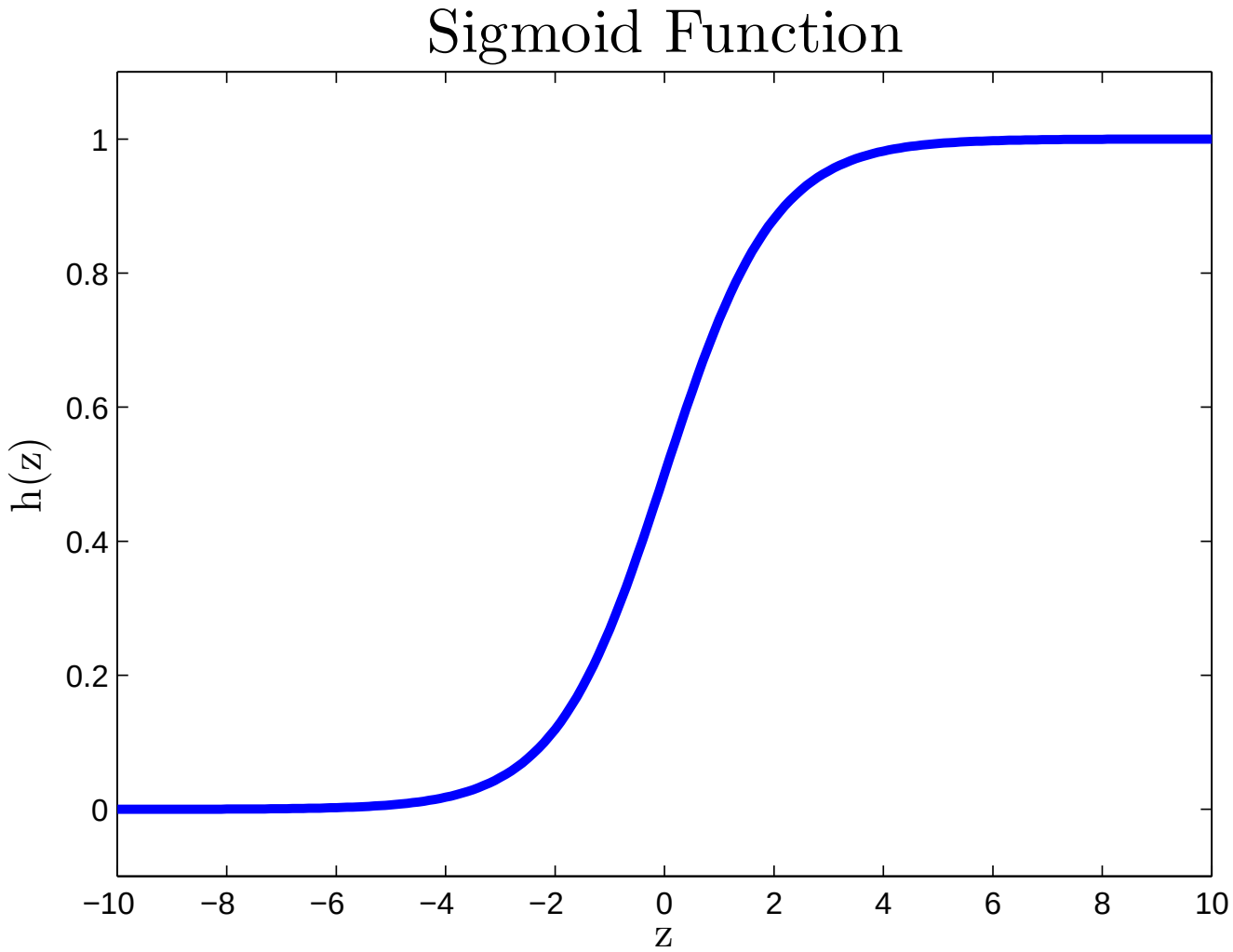


Рис. 1: *Сигмоид функция* $h(z) = \frac{1}{1+\exp(-z)}$

Сигмоид

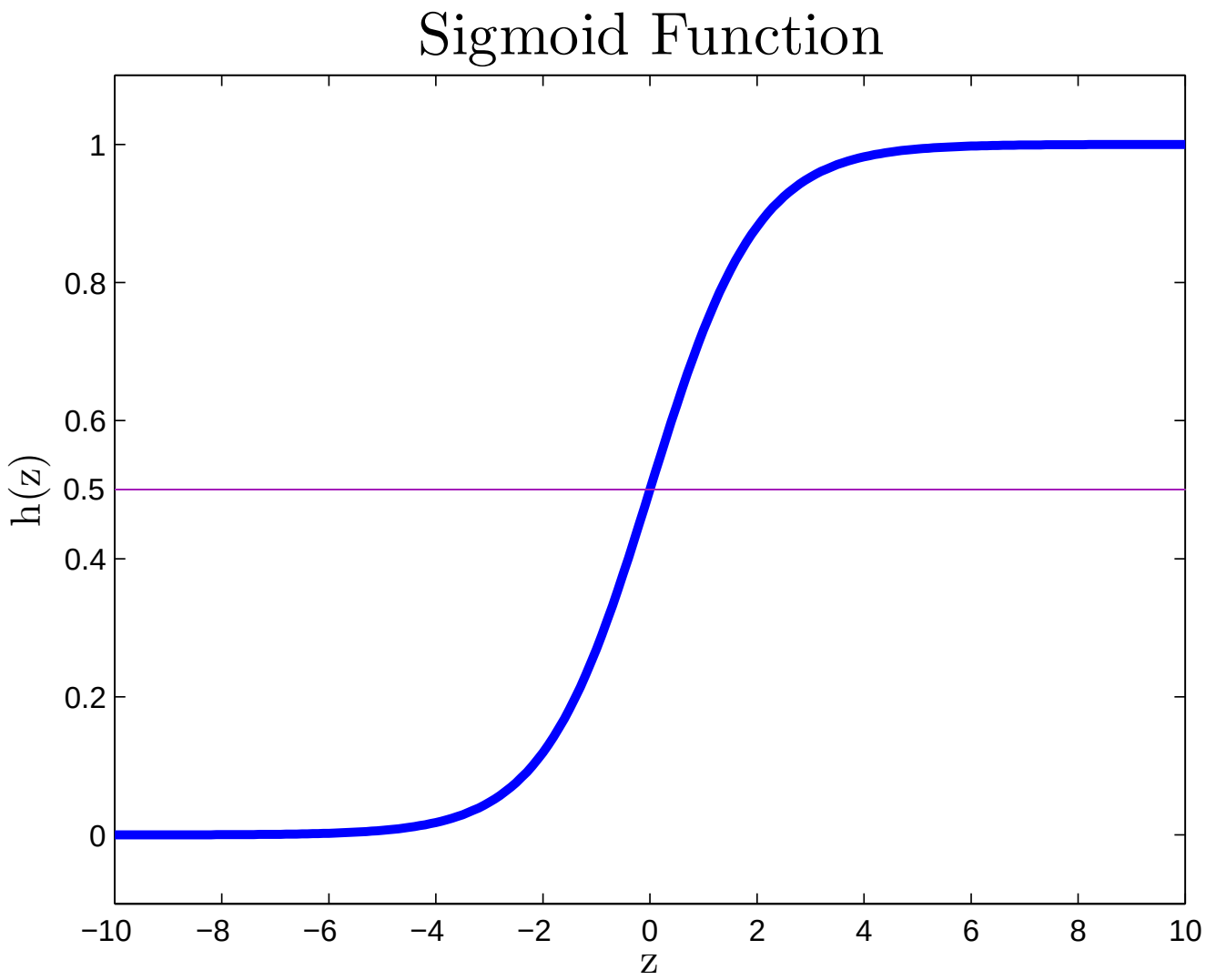


Рис. 1: *Сигмоид функция* $h(z) = \frac{1}{1+\exp(-z)}$

Решающая гиперплоскость

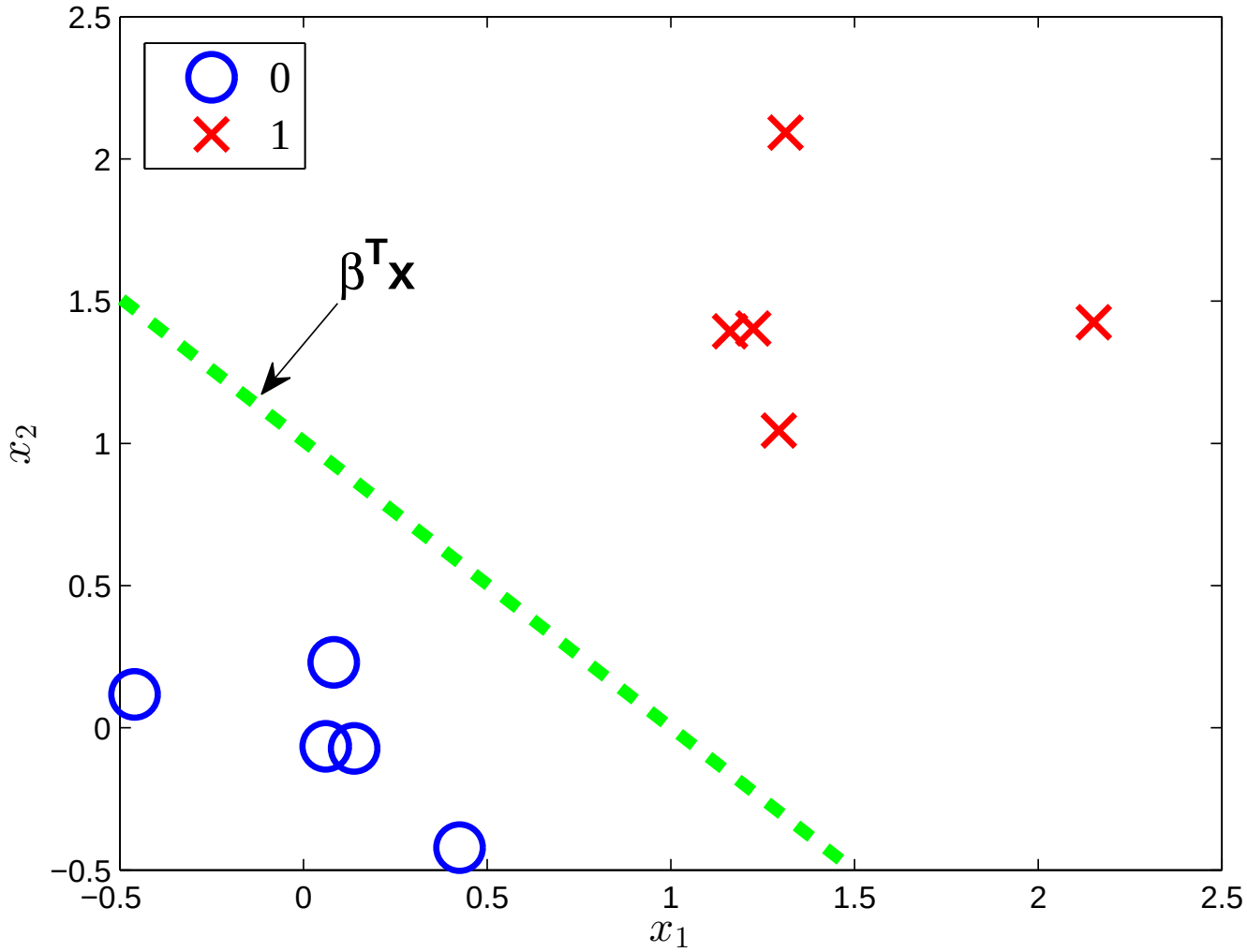


Рис. 2: *Решающая (разделяющая) гиперплоскость*

Поиск коэффициентов β – метод градиентного спуска

Функция потерь

Квадратичную функцию потерь $L(h(\beta, x), y(x)) = \sum_i (h(\beta, x_i) - y(x_i))^2$ использовать нецелесообразно, т. к. для логистической регрессии она не выпукла.

$$L(h(\beta, x), y(x)) = -\sum_i [y(x_i) \log(h(\beta, x_i)) + (1 - y(x_i)) \log(1 - h(\beta, x_i))]$$

Градиентный спуск

Итеративный метод поиска $\min_\beta L(h(\beta, x), y(x))$:

$$\beta_j = \beta_j - \alpha \frac{\partial}{\partial \beta_j} L(h(\beta, x), y(x)),$$

где α – скорость сходимости метода. Различные правила остановки.

Поиск коэффициентов β – метод градиентного спуска

Функция потерь

Квадратичную функцию потерь $L(h(\beta, x), y(x)) = \sum_i (h(\beta, x_i) - y(x_i))^2$ использовать нецелесообразно, т. к. для логистической регрессии она не выпукла.

$$L(h(\beta, x), y(x)) = -\sum_i [y(x_i) \log(h(\beta, x_i)) + (1 - y(x_i)) \log(1 - h(\beta, x_i))]$$

Градиентный спуск

Итеративный метод поиска $\min_\beta L(h(\beta, x), y(x))$:

$$\beta_j = \beta_j - \alpha \frac{\partial}{\partial \beta_j} L(h(\beta, x), y(x)),$$

где α – скорость сходимости метода. Различные правила остановки.

$$\frac{\partial}{\partial \beta_j} L(h(\beta, x), y(x)) = \sum_i (h(\beta, x_i) - y(x_i)) x_j$$

План лекции

1. Задача классификации
2. Метод 1-Rule
3. Метрические методы классификации
   - Метод k ближайших соседей
4. Байесовские методы классификации
   - Наивный байесовский классификатор (Naïve Bayes Classifier)
5. Логистическая регрессия
6. Меры качества классификации
   - Точность, полнота, F-мера
   - ROC-кривая и AUC

Точность, полнота, F-мера

Результаты некоторого алгоритма в задаче бинарной классификации можно описать в терминах матрицы ошибок (confusion matrix):

$$\begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix}, \text{где}$$

TP (True Positive) – число правильных предсказаний положительного класса, FP (False Positive) – ложных положительного, FN (False Negative) – ложных отрицательного класса, TN (True Negative) – верных отрицательного класса.

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

Необходимо найти баланс между точностью и полнотой, например, с помощью F-меры:

$$F_{\alpha,\beta}(Precision, Recall) = \frac{1}{\frac{1}{\alpha Precision} + \frac{1}{\beta Recall}}$$

Как быть в случае большего количества классов?

ROC-кривая (ROC — Reciever Operating Characteristic)

По оси X — False Positive Rate (доля ошибочно положительных классификаций)
По оси Y — True Positive Rate (доля правильных положительных классификаций)

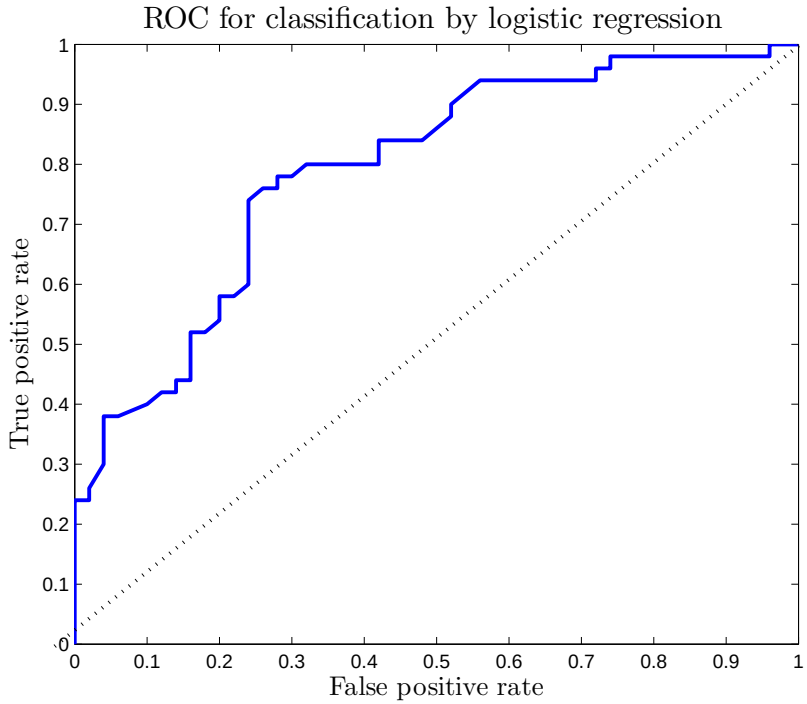


Рис. 3: *ROC-кривая для логистической регрессии*

Каждой точке ROC-кривой соответствует результат алгоритма при определенном значении параметра, например, значение порога вероятности (decision boundary).
Общая характеристика качества классификации определяется площадью под кривой — AUC (area under curve).

Вопросы и контакты

www.hse.ru/staff/dima

Спасибо!

dmitrii.ignatov[at]gmail.com