

Основы машинного обучения

Лекция 1.

Задачи машинного обучения. Обучение с
учителем. Метрические алгоритмы
классификации и регрессии

Москва, 2018

- ① Задача машинного обучения
- ② Данные
- ③ Обучение
- ④ Примеры задач
- ⑤ Метрические алгоритмы классификации

Задача машинного обучения. Обучение по прецедентам (с учителем)

X - множество объектов

Y - множество ответов

$y = f(x) : X \rightarrow Y$ - неизвестная зависимость

Дано:

Обучающая выборка $\{x_0 \dots x_l\} \subset X$

Выборка ответов $y_i = f(x_i), i = 1 \dots l$

Найти:

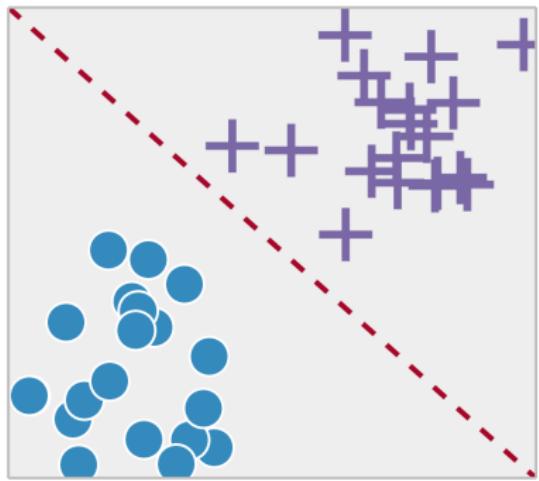
$f(x) - ?$

Задачи обучения с учителем

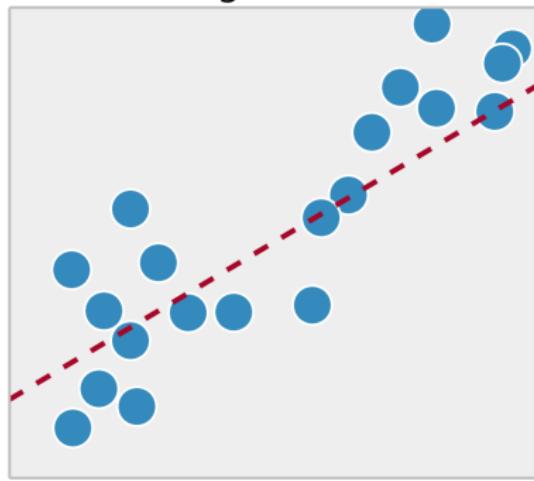
- Задача классификации
 - $Y = \{-1, 1\}$ - бинарная классификация
 - $Y = \{1 \dots M\}$ - задача классификации на M классов
 - $Y = \{0, 1\}^M$ - задача множественной классификации
- Задача восстановления регрессии
 - $Y \in \mathcal{R}$

Задачи обучения с учителем

Classification



Regression



$f_j : X \rightarrow D_j, j = 1 \dots n$ - признак

В зависимости от D_j определяются типы признаков:

- бинарный признак
- категориальный признак
- количественный признак

$(f_1(x), f_2(x), \dots, f_n(x))$ - вектор признаков для объекта x

Матрица "объект-признак" $\mathcal{F}_{l \times n}$

$$\begin{pmatrix} f_1(x_1) & f_2(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots & \dots \\ f_1(x_l) & f_2(x_l) & \dots & f_n(x_l) \end{pmatrix}$$

Функционалы ошибок

$\mathcal{L}(a, x)$ — функция потерь (loss function) — величина ошибки алгоритма $a \in A$ на объекте $x \in X$.

Функции потерь для задач классификации:

- $\mathcal{L}(a, x) = [a(x) \neq y(x)]$ — индикатор ошибки;

Функции потерь для задач регрессии:

- $\mathcal{L}(a, x) = |a(x) - y(x)|$ — абсолютное значение ошибки;
- $\mathcal{L}(a, x) = (a(x) - y(x))^2$ — квадратичная ошибка.

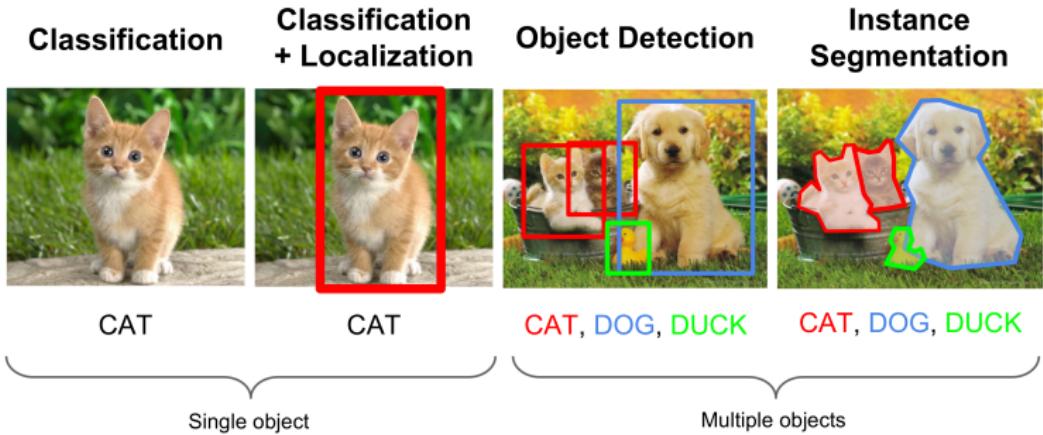
Эмпирический риск — функционал качества алгоритма a на X^ℓ :

$$Q(a, X^\ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} \mathcal{L}(a, x_i).$$

Пример данных. Табличные данные

Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	Education
41	Yes	Travel_Rarely	1102	Sales	1	2	Life Science
49	No	Travel_Frequently	279	Research & Development	8	1	Life Science
37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other
33	No	Travel_Frequently	1392	Research & Development	3	4	Life Science
27	No	Travel_Rarely	591	Research & Development	2	1	Medical
32	No	Travel_Frequently	1005	Research & Development	2	2	Life Science
59	No	Travel_Rarely	1324	Research & Development	3	3	Medical

Пример данных. Картинки



Пример данных. Текст

text	label
отвратительное обслуживание был у меня вклад в...	0
мнение о банке изменилось в худшую сторону это...	0
банк поступил красиво у меня дебетовая карта б...	1
прошу принять меры по исправлению ситуации бан...	0
спокойно и качественно пользуюсь услугами альф...	1

Модель машинного обучения

Модель машинного обучения (predictive model) - параметрическое семейство функций

$$A = \{g(X, \theta) | \theta \in \Theta\}$$

где $g : X \times \Theta \rightarrow Y$ - фиксированная функция,
 Θ - множество допустимых значений параметра θ

Примеры:

Линейная модель с вектором параметров $\theta = (\theta_0, \dots, \theta_n)$, $\Theta = \mathcal{R}^n$:

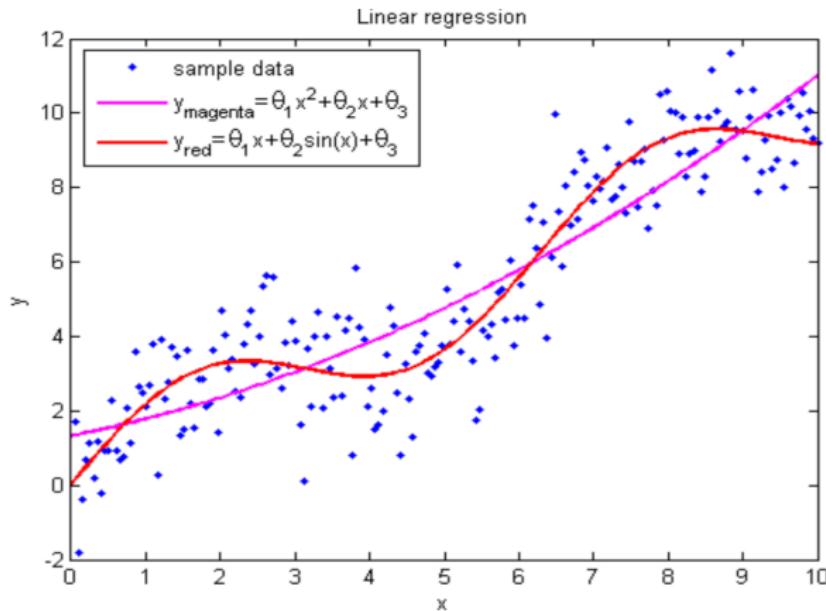
$g(x, \Theta) = \sum_{j=1}^n \Theta_j f_j(x)$ - для задачи регрессии, $Y = \mathcal{R}$

$g(x, \Theta) = a(\sum_{j=1}^n \Theta_j f_j(x))$ - для задачи классификации, a - функция активации

Например, $a(x) = sign(x)$ для $Y = \{-1, 1\}$, $a(x) = sigmoid(x) = \frac{1}{1+e^{-x}}$ для $Y \in [0, 1]$

Пример модели для задачи регрессии

$X = Y = \mathcal{R}$, $| = 200$, $n = 3$ признака: $\{1, x, x^2\}$
или $\{1, x, \sin(x)\}$



Пример. Нахождение редких распадов в физике

Данные - информация с детекторов

Классы - тип частицы (мюон, каон, пион, электрон, ...)

Примеры признаков

- **Количественные** - скорость, масса, угол, ...
- **Бинарные** - факт пролёта через разные слои детектора

Особенности задачи - модель должна быть физически-интерпретируемой

Пример. Кредитный скоринг

Данные - кредитные заявки

Классы - вернул/не вернул кредит

Примеры признаков

- **Количественные** - возраст, зарплата
- **Бинарные** - пол, факт наличия квартиры
- **Категориальные** - ВУЗ, место работы, марка машины

Особенности задачи - модель должна быть устойчива по времени

Пример. Классификация на больных-здоровых

Данные - информация о пациентах

Классы - типы болезней (здоров, ОРЗ, ОРВИ, грипп, ...)

Примеры признаков

- **Количественные** - показатели крови
- **Бинарные** - болел ли ветрянкой
- **Категориальные** - диагнозы, чем болел до этого

Особенности задачи - классов может быть много, один объект может принадлежать сразу разным классам

Пример. Оценка эмоциональной окрашенности текста

Данные - твиты/комментарии/отзывы

Классы - позитивная/негативная/(нейтральная)
эмоциональная окрашенность

Особенности задачи - текстовые данные
нужно как-то привести к числовым

Пример. Классификация картинок

Данные - картинки

Классы - на картинке кошка/собака/человек/...

Особенности задачи - как использовать
информацию о соседних пикселях?

Пример. Предсказание стоимости недвижимости

Данные - объявления

Целевая переменная - цена жилья

Примеры признаков

- **Количественные** - минут до метро, площадь, этаж
- **Бинарные** - балкон, ванная
- **Категориальные** - ближайшее метро, тип жилья

Особенности задачи:

- данные меняются со временем
- наличие категориальных признаков

Пример. Предсказание объёма продаж

Данные - тройка <товар, магазин, день>

Целевая переменная - объёмы продаж по товару и магазину

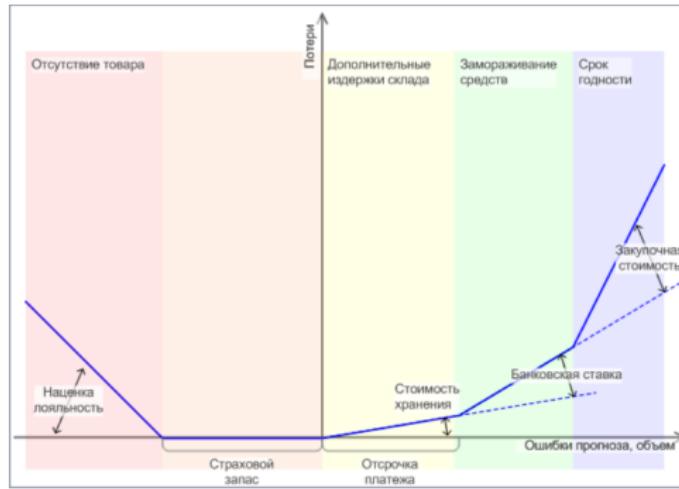
Примеры признаков

- **Количественные** - цена, вес, объём (стоимость хранения)
- **Бинарные** - выходной день, праздник, промоакция
- **Категориальные** - тип товара, местоположение магазина, ...

Особенности задачи

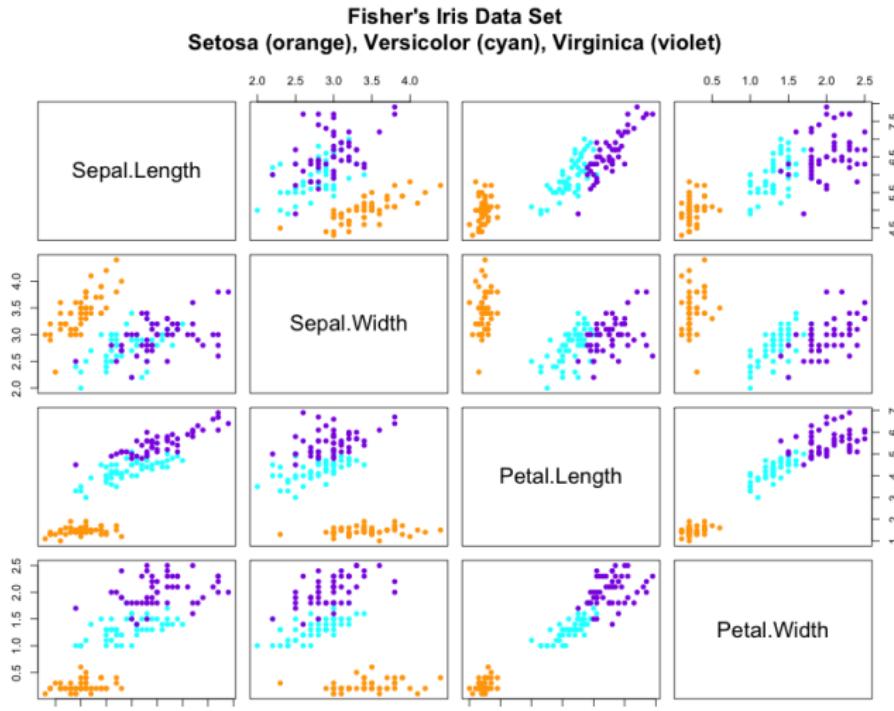
- сложная метрика качества

Пример. Предсказание объема продаж

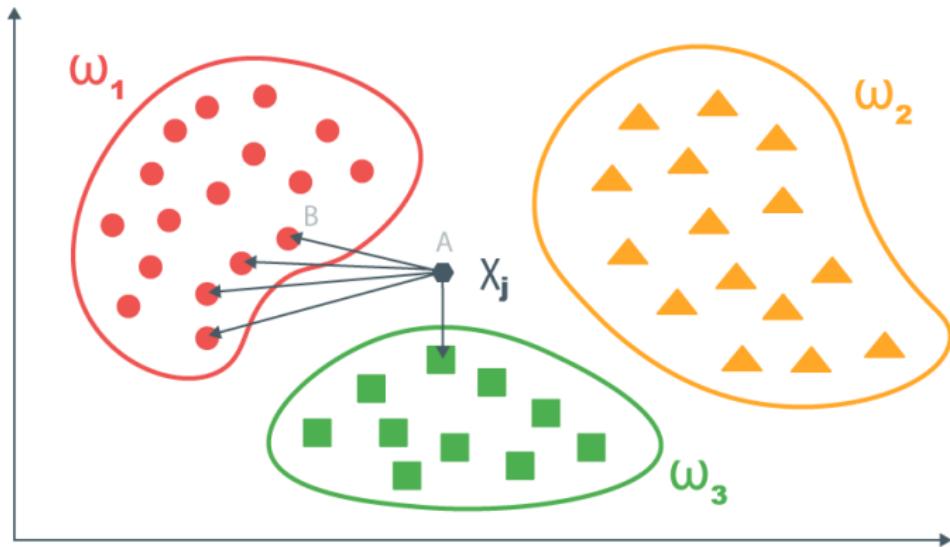


Метрические алгоритмы классификации.

Пример



Метод ближайших соседей



Метод ближайших соседей

Для произвольного $x \in X$ отранжируем объекты x_1, \dots, x_ℓ :

$$\rho(x, x^{(1)}) \leq \rho(x, x^{(2)}) \leq \dots \leq \rho(x, x^{(\ell)}),$$

$x^{(i)}$ — i -й сосед объекта x среди x_1, \dots, x_ℓ ;
 $y^{(i)}$ — ответ на i -м соседе объекта x .

Метрический алгоритм классификации:

$$a(x; X^\ell) = \arg \max_{y \in Y} \underbrace{\sum_{i=1}^{\ell} [y^{(i)} = y] w(i, x)}_{\Gamma_y(x)},$$

$w(i, x)$ — вес (степень важности) i -го соседа объекта x ,
неотрицателен, не возрастает по i .

$\Gamma_y(x)$ — оценка близости объекта x к классу y .

Метод окна Парзена

$w(i, x) = K\left(\frac{\rho(x, x^{(i)})}{h}\right)$, где h — ширина окна,
 $K(r)$ — ядро, не возрастает и положительно на $[0, 1]$.

Метод парзеновского окна *фиксированной ширины*:

$$a(x; X^\ell, \textcolor{red}{h}, K) = \arg \max_{y \in Y} \sum_{i=1}^{\ell} [y_i = y] K\left(\frac{\rho(x, x_i)}{\textcolor{red}{h}}\right)$$

Метод парзеновского окна *переменной ширины*:

$$a(x; X^\ell, \textcolor{red}{k}, K) = \arg \max_{y \in Y} \sum_{i=1}^{\ell} [y_i = y] K\left(\frac{\rho(x, x_i)}{\rho(x, x^{(k+1)})}\right)$$

Оптимизация параметров — по критерию LOO:

- выбор ширины окна h или числа соседей k
- выбор ядра K

Метод потенциальных функций

$$w(i, x) = \gamma^{(i)} K\left(\frac{\rho(x, x^{(i)})}{h^{(i)}}\right)$$

Более простая запись (здесь можно не ранжировать объекты):

$$a(x; X^\ell) = \arg \max_{y \in Y} \sum_{i=1}^{\ell} [y_i = y] \gamma_i K\left(\frac{\rho(x, x_i)}{h_i}\right),$$

где γ_i — веса объектов, $\gamma_i \geq 0$, $h_i > 0$.

Физическая аналогия из электростатики:

γ_i — величина «заряда» в точке x_i ;

h_i — «радиус действия» потенциала с центром в точке x_i ;

y_i — знак «заряда» (в случае двух классов $Y = \{-1, +1\}$);

$K(r) = \frac{1}{r}$ или $\frac{1}{r+a}$

В задачах классификации нет ограничений ни на K , ни на $|Y|$.

Формула непараметрической регрессии Надаля-Ватсона

Приближение константой $f(x, \alpha) = \alpha$ в окрестности $x \in X$:

$$Q(\alpha; X^\ell) = \sum_{i=1}^{\ell} w_i(x) (\alpha - y_i)^2 \rightarrow \min_{\alpha \in \mathbb{R}};$$

где $w_i(x) = K\left(\frac{\rho(x, x_i)}{h}\right)$ — веса объектов x_i относительно x ;
 $K(r)$ — ядро, невозрастающее, ограниченное, гладкое;
 h — ширина окна сглаживания.

Формула ядерного сглаживания Надаля–Ватсона:

$$a_h(x; X^\ell) = \frac{\sum_{i=1}^{\ell} y_i w_i(x)}{\sum_{i=1}^{\ell} w_i(x)} = \frac{\sum_{i=1}^{\ell} y_i K\left(\frac{\rho(x, x_i)}{h}\right)}{\sum_{i=1}^{\ell} K\left(\frac{\rho(x, x_i)}{h}\right)}.$$

Плюсы

- Наглядность и интерпретируемость
- Простота

Минусы

- Плохо работает на сложных задачах
- Требует хранения обучающей выборки в памяти
- Плохо работает на регрессии с трендом

Полезные ссылки

- <http://www.machinelearning.ru/wiki/images/f/fc/Voron-ML-Intro-slides.pdf> - Вводная лекция Воронцова
- <http://www.machinelearning.ru/wiki/images/c/c3/Voron-ML-Metric-slides.pdf> - Метрические методы классификации и регрессии Воронцова
- Видеолекции Воронцова на YouTube (Курс "Машинное обучение 2014")

Спасибо за внимание!