Обучение с учителем. Классификация. Функция риска. Логистическая регрессия. Feature selection и extraction

Третьякова Александра Леонидовна,
Волканова Маргарита Дмитриевна,
Федоров Никита Алексеевич

Санкт-Петербург
2019г.

Обучение с учителем. Постановка задачи

X — множество объектов, Y — множество ответов
$y : X \to Y$ — неизвестная зависимость (target function)

Дано: обучающая выборка — $(x_1, \ldots, x_n) \subset X$,
$y_i = y(x_i), \ i = 1, \ldots, n$ — известные ответы.
Найти: $a : X \to Y$ — функцию (decision function),
приближающую y на всем множестве X.

Вероятностная постановка задачи: имеется неизвестное распределение на множестве $X \times Y$ с плотностью $p(x, y)$, из которого случайно выбираются $\mathbb{X}_n = (x_i, y_i)_{i=1}^n$ (независимые).

Задача классификации:

- $Y = \{-1, +1\}$ — классификация на 2 класса
- $Y = \{1, \ldots, K\}$ — классификация на K классов

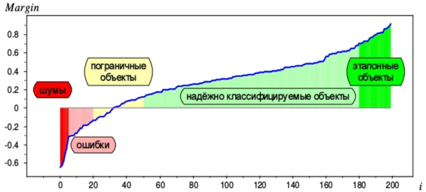
Обучение с учителем. Постановка задачи

Обучающая выборка: $\mathbb{X}_n = (x_i, y_i)_{i=1}^n$, $x_i \in \mathbb{R}^p$, $y_i \in \{-1, +1\}$.
$f(x, w)$ — разделяющая (дискриминантная) функция, $w \in \mathbb{R}^p$.
$a(x, w) = \operatorname{sign} f(x, w)$ — классификатор.
$f(x, w) = 0$ — разделяющая поверхность.

$M_i(w) = y_i f(x_i, w)$ — отступ объекта x_i.
Если $M_i(w) < 0$, то классификатор ошибается на x_i.

Задача

$$Q(w) = \sum_{i=1}^{n} [M_i(w) < 0] \leq \tilde{Q}(w) = \sum_{i=1}^{n} \mathcal{L}(M_i(w)) \to \min_{w}$$



Примеры функции потерь

- Пороговая функция потерь: $[M_i(w) < 0]$
- Логарифмическая: $\log_2(1 + e^{-M_i(w)})$
- Экспоненциальная: $e^{-M_i(w)}$
- Кусочно-линейная: $(1 - M_i(w))_+$
- Квадратичная: $(1 - M_i(w))^2$

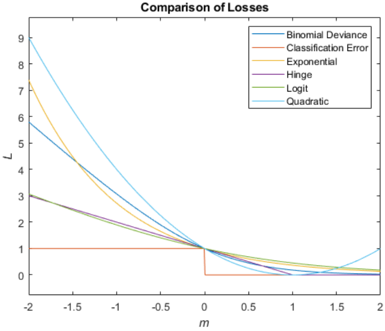


Рис.: Примеры функции потерь $\mathcal{L}(M)$

Линейный классификатор

$f_j : X \to \mathbb{R},\ \ j = 1, \dots, p$

Линейная модель классификации:

$$a(x, w) = \mathrm{sign}\Big(\sum_{j=1}^{p} w_j f_j(x) - w_0\Big), \quad w_0, w_1, \dots, w_p \in \mathbb{R}.$$

Пусть $f_0 = -1$, тогда

$$a(x, w) = \mathrm{sign}\langle w, x\rangle,\ \ x, w \in \mathbb{R}^{p+1}.$$

$M_i(x) = y_i\langle w, x_i\rangle$ — отступ объекта x_i.

Задача

$$Q(w) = \sum_{i=1}^{n}[y_i\langle w, x_i\rangle < 0] \le \sum_{i=1}^{n} \mathcal{L}(y_i\langle w, x_i\rangle) \to \min_{w}$$

Проверка по тестовой выборке $\tilde{\mathbb{X}}_k = (\tilde{x}_i, \tilde{y}_i)_{i=1}^{k}$:

$$\bar{Q}(w) = \frac{1}{k}\sum_{i=1}^{k}[\tilde{y}_i\langle w, \tilde{x}_i\rangle < 0]$$

Решение задачи оптимизации

Для решения задачи оптимизации используется **метод стохастического градиента**.

Вход: $\mathbb{X}_n = (x_i, y_i)_{i=1}^n, h, \lambda$
Выход: w

1. Инициализация $w_j, \ j = 0, \ldots, p$
2. Инициализация $\bar{Q}(w)$
3. Повторять
   1. Выбор x_i из $\mathbb{X}_n$ случайным образом
   2. Вычисление $\varepsilon_i := \mathcal{L}_i(w)$
   3. Градиентный шаг $w := w - h\nabla\mathcal{L}_i(w)$
   4. Вычисление $\bar{Q}(w) := \lambda\varepsilon_i + (1-\lambda)\bar{Q}(w)$

   пока $\bar{Q}(w)$ или w не сойдутся

Регуляризация

Проблемы:

- Признаков намного больше, чем объектов
- Мультиколлинеарность признаков:
  Пусть
  $a(x, w) = \mathrm{sign}(w_1 f_1(x) + w_2 f_2(x) - w_0), \quad f_2(x) = k f_1(x).$
  Тогда
  $w_1 f_1(x) + w_2 f_2(x) = (w_1 + \beta) f_1(x) + (w_2 - k\beta) f_2(x) \; \forall \beta$

Таким образом, очень много различных векторов дадут близкие значения функционала качества, но при этом коэффициенты могут существенно отличаться. Признаком такого явления может являться большая $\|w\|$.

Задача с регуляризацией

$$Q(w, \mathbb{X}_n) = Q(w, \mathbb{X}_n) + \frac{\tau}{2} \|w\|^2 \to \min_w$$

Связь с принципом максимума правдоподобия

Рассмотрим модель: пусть $X \times Y$ — вероятностное пространство с плотностью $p(x, y|w)$.
Пусть $\mathbb{X}_n = (x_i, y_i)_{i=1}^n \sim p(x, y|w)$ — независимы, одинаково распределены.

- Максимизация правдоподобия:
  $$L(w; \mathbb{X}_n) = \ln \prod_{i=1}^{n} p(x_i, y_i|w) = \sum_{i=1}^{n} \ln p(x_i, y_i|w) \to \max_w.$$
- Минимизация аппроксимированного эмпирического риска:
  $$\tilde{Q}(w; \mathbb{X}_n) = \sum_{i=1}^{n} \mathcal{L}(y_i f(x_i, w)) \to \min_w.$$

Эти задачи эквивалентны, если положить $-\ln p(x_i, y_i|w) = \mathcal{L}(y_i f(x_i, w))$.

Пример: $p(x, y|w) = \frac{1}{1+\exp(-y\langle x,w\rangle)}$ (сигмоидная функция), $\mathcal{L}(yf(x, w)) = \log(1 + \exp(-y\langle x, w\rangle))$ (логарифмическая функция потерь)

Модель: $\mathbb{X}_n = (x_i, y_i)_{i=1}^n \sim p(x, y|w)$ — н.о.р.,
Пусть $w \sim p(w; \gamma)$, γ — вектор гиперпараметров. Тогда:

$$p(\mathbb{X}_n, w) = p(\mathbb{X}_n|w)p(w; \gamma)$$

Принцип максимума правдоподобия:

$$L(w; \mathbb{X}_n) = \ln p(\mathbb{X}_n, w) = \sum_{i=1}^{n} p(x_i, y_i|w) + \underbrace{\ln p(w; \gamma)}_{\text{регуляризатор}} \to \max_{w,\gamma}$$

Примеры:

1. Гауссовский регуляризатор: $p(w; \sigma) = \frac{1}{(2\pi\sigma)^{p/2}} \exp -\frac{\|w\|^2}{2\sigma}$,
   тогда $-\ln p(w; \sigma) = \frac{1}{2\sigma}\|w\|^2 + \text{const} \quad (\tau = 1/\sigma)$
2. Регуляризатор Лапласа (приводит к отбору признаков):
   $p(w; C) = \frac{1}{(2C)^p} \exp -\frac{\|w\|_1}{C}$,
   тогда $-\ln p(w; C) = \frac{1}{C} \sum_{j=1}^{p} |w_j| + \text{const} \quad (\tau = 1/C)$

Регуляризатор Лапласа. Отбор признаков

Задача: $Q(w; \mathbb{X}_n) = \sum\limits_{i=1}^{n} \ln p(x_i, y_i | w) + \frac{1}{C} \sum\limits_{j=1}^{p} |w_j| \to \min\limits_{w,C}.$

Замена: $\begin{cases} u_j = \frac{1}{2}(|w_j| + w_j) \\ v_j = \frac{1}{2}(|w_j| - w_j) \end{cases}$, тогда $\begin{cases} w_j = u_j - v_j \\ |w_j| = u_j + v_j \end{cases}$,

$$\begin{cases} Q(u, v) = \sum\limits_{i=1}^{n} \mathcal{L}(M_i(u - v, w_0)) + \frac{1}{C} \sum\limits_{j=1}^{p} (u_j + v_j) \to \min\limits_{u,v} \\ u_j \geq 0, \ v_j \geq 0, \quad j = 1, \dots, p \end{cases}$$

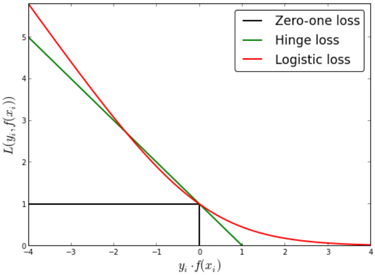
При уменьшении C (возрастании $\frac{1}{C}$) обнуляются u_j и v_j для все большего количества j, то есть $w_j = 0$ и признак не учитывается.

При $C \to 0$ выбросим все признаки.

Логистическая регрессия. Подход через минимизацию функции потерь

Линейная модель классификации:
$a(x) = \mathrm{sign}\langle w, x\rangle, \quad x, w \in \mathbb{R}^p, \quad M = \langle w, x\rangle y$ — отступ.

В качестве аппроксимации пороговой функции потерь берется логарифмическая функция потерь $\mathcal{L}(M) = \log(1 + e^{-M})$.



Логистическая регрессия. Подход через минимизацию функции потерь

Задача

$$Q(w) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \langle w, x_i \rangle)) \to \min_{w}$$

Методы решения задачи минимизации:

- метод стохастического градиента
- метод Ньютона-Рафсона

$P(y|x,w) = \sigma_w(M) = \frac{1}{1+e^{-\langle x,w\rangle y}}$ — сигмоидная функция.

Свойства $\sigma(z)$:

- $\sigma(z) \in [0,1]$, задана на $(-\infty, +\infty)$
- $\sigma(z) \to 1,\ z \to +\infty;$
  $\sigma(z) \to 0,\ z \to -\infty$
- $\sigma(z) + \sigma(-z) = 1$
- $\sigma'(z) = \sigma(z)\sigma(-z)$

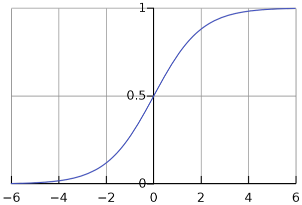


Рис.: Сигмоидная функция

Логистическая регрессия. Вероятностный подход

Пусть $Y = \{0, 1\}$.

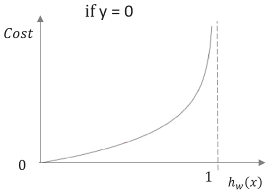
- $P(y_i = 1 | x; w) = \sigma_w(x)$
- $P(y_i = 0 | x; w) = 1 - \sigma_w(x)$

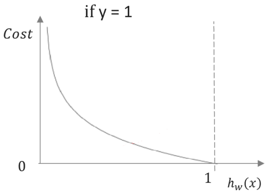
Тогда $P(y|x; w) = (\sigma_w(x))^y (1 - \sigma_w(x))^{1-y}$.

Функция правдоподобия:

$$Q(w) = -\log L(w) = -\log \prod_{i=1}^{n} (\sigma_w(x_i))^{y_i} (1 - \sigma_w(x_i))^{1-y_i} =$$

$$= -\sum_{i=1}^{n} [y_i \log(\sigma_w(x_i) + (1 - y_i)) \log(1 - \sigma_w(x_i))] \to \min_w$$





Линейная и логистическая регрессия

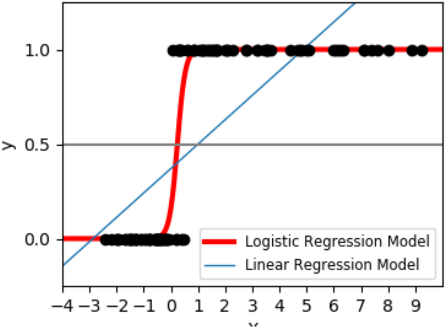


Рис.: Линейная и логистическая регрессия

Логистическая регрессия. Регуляризация

$$Q(w) = -\sum_{i=1}^{n}[y_i \log(\sigma_w(x_i) + (1 - y_i)) \log(1 - \sigma_w(x_i))]$$

Регуляризация в логистической регрессии:

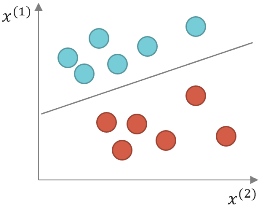
- L2: $Q_\tau(w) = Q(w) + \frac{\tau}{2}\sum_{j=1}^{p} w_j^2 \to \min\limits_w$
- L1: $Q_\tau(w) = Q(w) + \tau\sum_{j=1}^{p} |w_j| \to \min\limits_w$

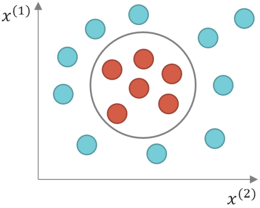
Параметр τ можно подбирать с помощью кросс-валидации.

Методы решения задачи минимизации: метод стохастического градиента, метод Ньютона-Рафсона.

Логистическая регрессия. Добавление нелинейных признаков

Можно ли использовать логистическую регрессию в случае, когда нет линейной разделимости?





$\langle w, x \rangle = w_1 x_1 + w_2 x_2 + w_0$

Разделяющая поверхность:

$w_1 x_1 + w_2 x_2 + w_0 = 0$

$\langle w, x \rangle = w_1 x_1^2 + w_2 x_2^2 + w_0$

Разделяющая поверхность:

$w_1 x_1^2 + w_2 x_2^2 + w_0 = 0$

Многоклассовая логистическая регрессия

Линейный классификатор при произвольном числе классов $Y = \{1, \dots, K\}$:

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

Вероятность того, что объект x относится к классу i:

$$P(y = i | x; w) = \frac{\exp \langle w_y, x \rangle}{\sum\limits_{z \in Y} \exp \langle w_z, x \rangle} = \frac{e^{w_i^T x}}{\sum\limits_{k=1}^{K} e^{w_k^T x}}$$

Задача:

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

Логистическая регрессия. Преимущества и недостатки

Плюсы:

1. Позволяет оценить вероятности принадлежности объектов к классу
2. Достаточно быстро работает при больших объемах выборки
3. Применима в случае отсутствия линейной разделимости, если на вход подать полиномиальные признаки

Минусы:

1. Плохо работает в задачах, в которых зависимость сложная, нелинейная

Пример использования логистической регрессии. Задача кредитного скоринга

Пусть $Y = \{+1, -1\}$

Величина потери $D_{xy} = \begin{cases} S(x), & Y = -1 \text{ (кредит не вернули)} \\ -rS(x), & Y = +1 \text{ (кредит вернули)} \end{cases}$.

Логистическая регрессия дает возможность вычислять апостериорные вероятности принадлежности классу для каждого объекта x:

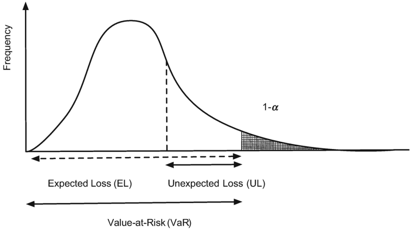
$$P(y|x;w) = \frac{1}{1 + e^{-\langle x,w\rangle y}}$$

$R(x) = \sum_{y \in Y} D_{xy} P(y|x) = \sum_{y \in Y} D_{xy} \sigma_w(x)$ — оценка мат. ожидания потерь для объекта x. Хотим узнать, сколько банк потеряет в худшем случае.

Пример использования логистической регрессии. Задача кредитного скоринга

Строим эмпирическую функцию распределения потерь.
Метод Value at Risk:

1. N раз ($N = 1000$):
   - $\forall x_i$ случайно разыгрываем $y_i \sim P(y|x_i), \quad i = 1, \ldots, n$
   - вычисляем суммарные потери $V = \sum_{i=1}^{n} D_{x_i y_i}$
2. строим эмпирическое распределение величины V
3. 99%-квантиль показывает величину резервируемого капитала



Выводы

- Существуют различные варианты аппроксимации пороговой функции потерь, позволяющие использовать методы градиентной оптимизации
- Регуляризация решает проблему мультиколлинеарности
- Минимизация аппроксимированного эмпирического риска и максимизация правдоподобия оказываются эквивалентными задачами
- Логистическая регрессия позволяет оценить условные вероятности классов
- В случае отсутствия линейной разделимости можно добавить нелинейные признаки и использовать логистическую регрессию