Санкт-Петербургский государственный университет
Прикладная математика и информатика
Статистическое моделирование

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

Третьякова Александра Леонидовна
Волканова Маргарита Дмитриевна
Федоров Никита Алексеевич

Санкт-Петербург
2020

Оглавление

1. Обучение с учителем. Постановка задачи

Пусть наблюдается некоторый количественный отклик Y и p предикторов (признаков) $X_1, \ldots, X_p$. Будем предполагать, что между Y и $X = (X_1, \ldots, X_p)$ существует определенная связь, которую можно представить в виде

$$Y = f^*(X) + \varepsilon,$$

где f^* — фиксированная, но неизвестная функция от предикторов, ε — ошибка, которая не зависит от X и имеет нулевое среднее значение.

Так как ошибки имеют нулевое среднее, можем предсказывать Y в соответствии с формулой

$$\hat{Y} = \hat{f}(X, \theta),$$

где $\hat{f}$ — оценка f^*, θ — вектор параметров оценки, $\hat{Y}$ — предсказанное значение Y.

Когда целью является предсказание, нам не важна точная форма функции $\hat{f}$, если она обеспечивает точные предсказания Y.

Пусть имеется выборка из n отдельных наблюдений: $x_1, \ldots, x_n$ для которых известны соответствующие значения отклика. Обозначим x_{ij} — значение j-го признака i-го наблюдения, $x_i = (x_{i1}, \ldots, x_{ip})^\mathsf{T}$, y_i — отклик у i-го наблюдения.

Хотим найти такую функцию $\hat{f}$, что $y \approx \hat{f}(x)$ для любого наблюдения (x, y). Совокупность X_n пар $(x_1, y_1), \ldots, (x_n, y_n)$, которая участвует в оценке функции f^*, называется обучающей выборкой. Выборка $X'_k = (x'_i, y'_i)_{i=1}^k$, не участвующая в оценке функции f^*, называется тестовой (или контрольной).

Вероятностная постановка задачи: $Y \sim Ber(\sigma(X))$ при классификации на два класса и $Y \sim Mult(\sigma(X))$, где $\sigma(X) = (\sigma_1(X), \ldots, \sigma_K(X))$ при классификации на K классов. Задача — построить оценку $\hat{\sigma} = \hat{\sigma}(X, \theta)$ параметра σ. Предсказание $\hat{Y} = \operatorname{argmax}_k \sigma_k(X)$.

2. Функция риска

Для оценки качества предсказания необходимо ввести функционал качества. Для этого сначала введём функцию потерь.

Определение 1. *Функция потерь (loss function) $L((x, y), \theta)$ — неотрицательная функция, характеризующая величину ошибки предсказания на объекте x.*

Для задачи классификации в качестве функции потерь обычно используется индикатор ошибки $L((x,y),\theta) = \delta(y, \hat{f}(x,\theta))$.

Теперь определим эмпирический риск — функционал, который используется для оценки качества предсказания.

Определение 2. *Эмпирический риск* $Q(X_n,\theta) = \dfrac{1}{n}\sum\limits_{i=1}^{n} L((x_i,y_i),\theta)$

3. Регуляризация

Проблемы:

- Признаков намного больше, чем объектов
- Мультиколлинеарность признаков:
  Пусть $\hat{f}(x,\theta) = \text{sign}(\theta_1 f_1(x) + \theta_2 f_2(x) - \theta_0)$, $f_2(x) = k f_1(x)$.
  Тогда $\theta_1 f_1(x) + \theta_2 f_2(x) = (\theta_1 + \beta) f_1(x) + (\theta_2 - k\beta) f_2(x)\ \forall \beta$

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

Задача с регуляризацией:

$$\widetilde{Q}(X_n,\theta) = Q(X_n,\theta) + \frac{\tau}{2}\|\theta\|^2 \to \min_{\theta}$$

4. Связь с принципом максимума правдоподобия

Рассмотрим **модель**: пусть Y — вероятностное пространство с плотностью $p(y|x,\theta)$, то есть $\mathcal{L}(Y)$ — семейство распределений, зависящее от параметров X и θ.

Пусть $X_n = (x_i,y_i)_{i=1}^n$, и y_i — независимы, одинаково распределены.

- **Максимизация правдоподобия:**

$$\mathsf{L}(\theta; X_n) = \ln \prod_{i=1}^{n} p(y_i|x_i,\theta) = \sum_{i=1}^{n} \ln p(y_i|x_i,\theta) \to \max_{\theta}.$$

- **Минимизация аппроксимированного эмпирического риска:**

$$\tilde{Q}(X_n,\theta) = \sum_{i=1}^{n} L(x_i,y_i,\theta) \to \min_{\theta}.$$

Эти задачи эквивалентны, если положить $-\ln p(y_i|x_i,\theta) = L(x_i,y_i,\theta)$.

Пример:

- $p(y|x,\theta) = \frac{1}{1+\exp(-y\langle x,\theta\rangle)}$ — сигмоидная функция.
- $L(x,y,\theta) = \log(1+exp(-y\langle x,\theta\rangle))$ — логарифмическая функция потерь.

Принцип максимума правдоподобия: Пусть $\theta \sim p(\theta;\gamma)$.

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

Примеры:

1. **Гауссовский регуляризатор**:

$$\mathsf{p}(\theta;\sigma) = \frac{1}{(2\pi\sigma)^{\mathsf{p}/2}} \exp -\frac{\|\theta\|^2}{2\sigma},$$

тогда

$$-\ln \mathsf{p}(\theta;\sigma) = \frac{1}{2\sigma}\|\theta\|^2 + \text{const} \quad (\tau = 1/\sigma)$$

2. **Регуляризатор Лапласа** (приводит к отбору признаков):

$$\mathsf{p}(\theta;\mathsf{C}) = \frac{1}{(2\mathsf{C})^{\mathsf{p}}} \exp -\frac{\|\theta\|_1}{\mathsf{C}},$$

тогда

$$-\ln \mathsf{p}(\theta;\mathsf{C}) = \frac{1}{\mathsf{C}} \sum_{\mathsf{j}=1}^{\mathsf{p}} |\theta_{\mathsf{j}}| + \text{const} \quad (\tau = 1/\mathsf{C})$$

5. Регуляризатор Лапласа. Отбор признаков

Задача:

$$Q(X_n,\theta) = \sum_{i=1}^{n} \ln p(y_i|x_i,\theta) + \frac{1}{C}\sum_{j=1}^{p} |\theta_j| \to \min_{\theta,C}.$$

Замена:

$$\begin{cases} \mathsf{u}_{\mathsf{j}} = \frac{1}{2}(|\theta_{\mathsf{j}}| + \theta_{\mathsf{j}}) \\ \mathsf{v}_{\mathsf{j}} = \frac{1}{2}(|\theta_{\mathsf{j}}| - \theta_{\mathsf{j}}) \end{cases},$$

тогда

$$\begin{cases} \theta_{\mathsf{j}} = \mathsf{u}_{\mathsf{j}} - \mathsf{v}_{\mathsf{j}} \\ |\theta_{\mathsf{j}}| = \mathsf{u}_{\mathsf{j}} + \mathsf{v}_{\mathsf{j}} \end{cases}$$

$$\begin{cases} Q(u,v) = \sum\limits_{i=1}^{n} \mathcal{L}(M_i(u-v,\theta_0)) + \frac{1}{C}\sum\limits_{j=1}^{p}(u_j+v_j) \to \min\limits_{u,v} \\ u_j \geq 0,\ v_j \geq 0, \quad j = 1,\ldots,p \end{cases}$$

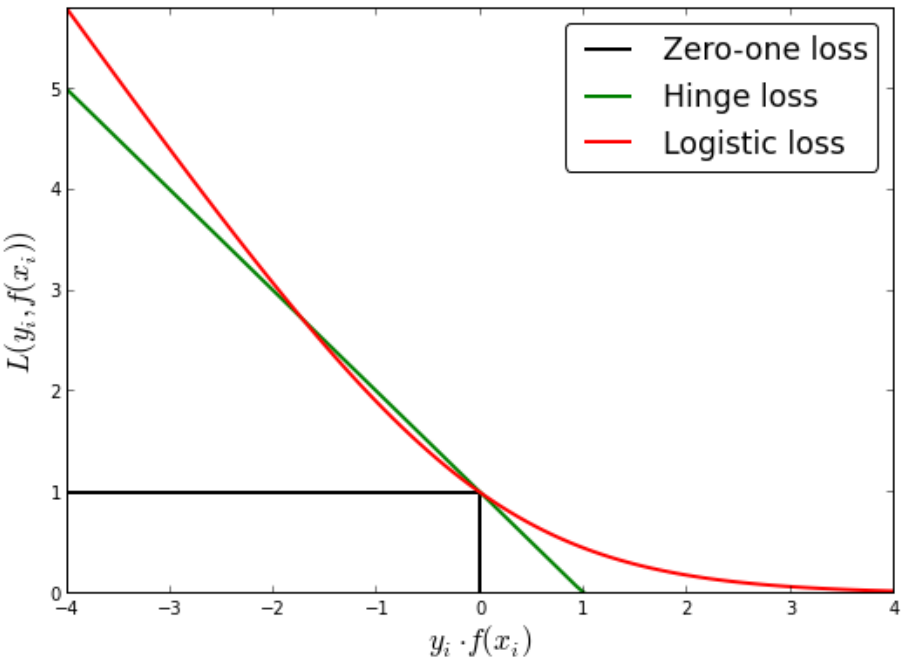
При уменьшении C (возрастании $\frac{1}{C}$) обнуляются u_j и v_j для все большего количества j, то есть $\theta_j = 0$ и признак не учитывается. При $C \to 0$ выбросим все признаки.

6. Логистическая регрессия. Подход через минимизацию функции потерь

Линейная модель классификации:

- $\hat{f}(x) = \text{sign}\langle\theta, x\rangle, \quad x, \theta \in \mathbb{R}^p$
- $M = \langle\theta, x\rangle y$ — отступ.

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



Задача 1. $Q(X_n, \theta) = \sum\limits_{i=1}^{n} \log(1 + \exp(-y_i\langle\theta, x_i\rangle)) \to \min\limits_{\theta}$

Методы решения задачи минимизации:

- метод стохастического градиента
- метод Ньютона-Рафсона

7. Логистическая регрессия. Вероятностный подход

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

Свойства $\sigma(\mathsf{z})$:

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

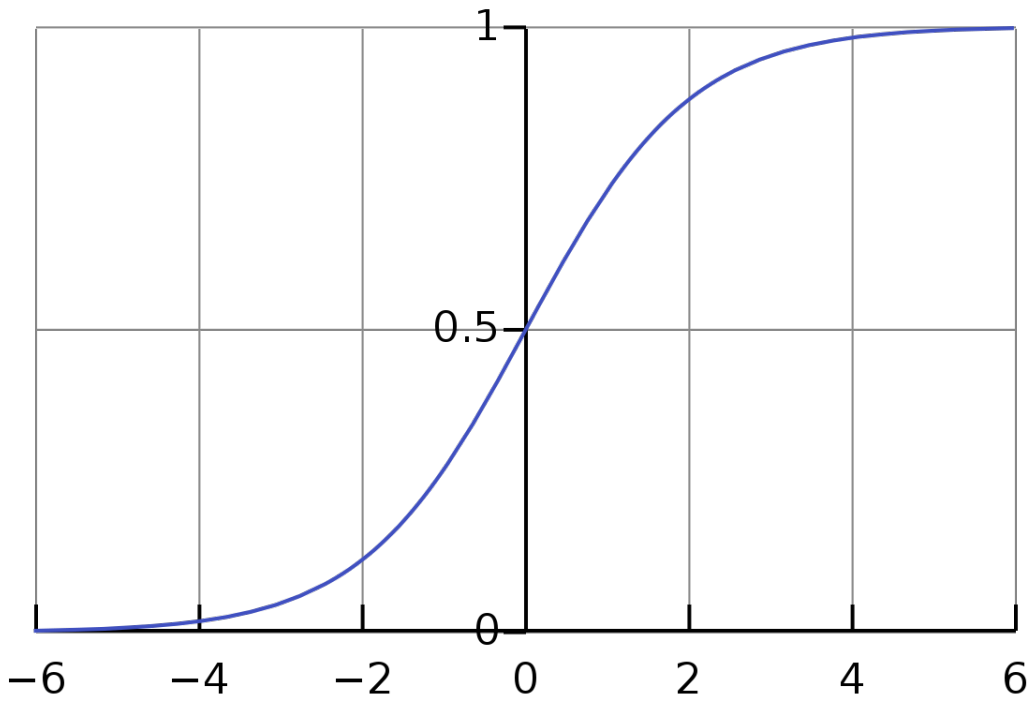


Рис. 1. Сигмоидная функция

Пусть $\mathsf{Y} = \{0,1\}$.

- $\mathsf{P}(\mathsf{y_i} = 1|\mathsf{x};\theta) = \sigma_\theta(\mathsf{x})$

- $P(y_i = 0|x;\theta) = 1 - \sigma_\theta(x)$

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

Функция правдоподобия:

$$Q(X_n, \theta) = -\log L(\theta) = -\log \prod_{i=1}^{n} (\sigma_\theta(x_i))^{y_i}(1 - \sigma_\theta(x_i))^{1-y_i} =$$

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

8. Линейная и логистическая регрессия

Существуют примеры данных, для которых логистическая регрессия показывает лучшие результаты, чем линейная.

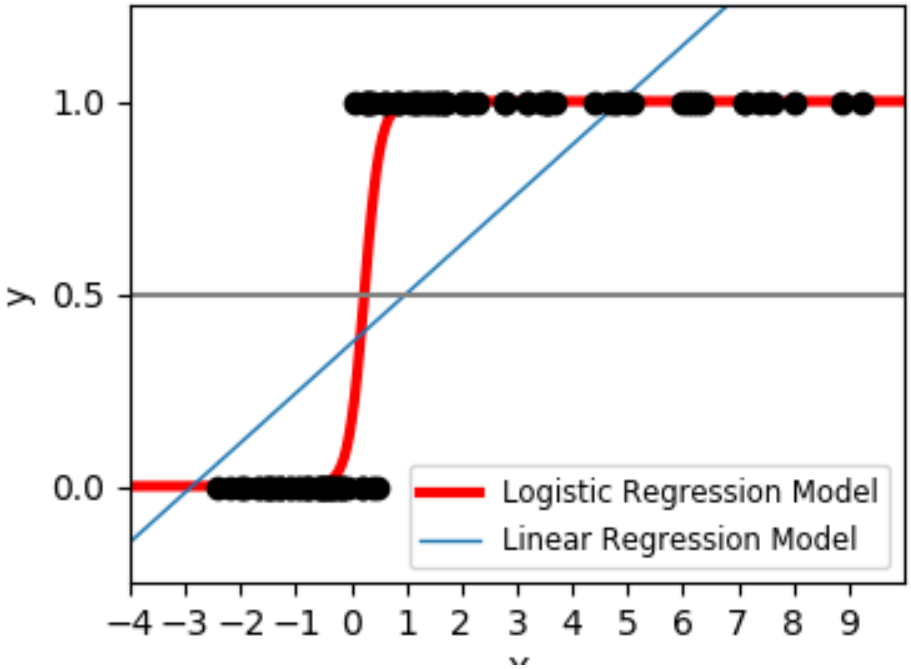


Рис. 2. Линейная и логистическая регрессия

9. Логистическая регрессия. Регуляризация

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

Регуляризация в логистической регрессии:

- **L2**: $\mathsf{Q}_\tau(\theta) = \mathsf{Q}(\theta) + \frac{\tau}{2}\sum_{\mathsf{j}=1}^{\mathsf{p}}\theta_\mathsf{j}^2 \to \min_\theta$
- **L1**: $\mathsf{Q}_\tau(\theta) = \mathsf{Q}(\theta) + \tau\sum_{\mathsf{j}=1}^{\mathsf{p}}|\theta_\mathsf{j}| \to \min_\theta$

Параметр τ можно подбирать с помощью кросс-валидации.

Методы решения задачи минимизации:

- метод стохастического градиента
- метод Ньютона-Рафсона.

10. Многоклассовая логистическая регрессия

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

$$\hat{f}(x, \theta) = \arg\max_{y \in Y}\langle\theta_y, x\rangle,\ \ x, \theta_y \in \mathbb{R}^p$$

Вероятность того, что объект x относится к классу i:

$$\mathsf{P}(y = i|x;\theta) = \frac{\exp\langle\theta_y, x\rangle}{\sum\limits_{z\in Y}\exp\langle\theta_z, x\rangle} = \frac{e^{\theta_i^{\mathrm{T}}x}}{\sum\limits_{k=1}^{K}e^{\theta_k^{\mathrm{T}}x}}$$

Задача:

$$\mathsf{Q}(\mathsf{X_n}, \theta) = -\sum_{\mathsf{i}=1}^{\mathsf{n}}\log\mathsf{P}(\mathsf{y_i}|\mathsf{x_i};\theta) \to \min_\theta$$

11. 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

12. 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