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

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

где f — вид функции, которой приближаем зависимость f^*, $\hat{\theta}$ — вектор параметров оценки, $\hat{Y}$ — предсказанное значение Y.

Когда целью является предсказание, нам не важна точная форма функции $f(X, \hat{\theta})$, 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 f, что $y \approx f(x)$ 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 $\hat{Y} = \mathrm{argmax}_k\ \sigma_k(X)$.

2. Линейный классификатор

Пусть $f_j : X \to \mathbb{R},\ j = 1, \ldots, p$. Линейная модель классификации:

$$f(x, \theta) = \mathrm{sign}(\theta_0 + \sum_{j=1}^{p} \theta_j f_j(x)).$$

Если добавить к признакам единицу и положить $f_j(x) = x$, получаем

$$f(x, \theta) = \operatorname{sign}\langle x, \theta\rangle.$$

3. Функция риска

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

Определение 1. *Отступ* $\mathsf{M} = \langle\theta, \mathsf{x}\rangle\mathsf{y}$ *— «расстояние» между реальным значением и предсказанным.*

Определение 2. *Функция потерь (loss function) $L(M)$ — неотрицательная функция, характеризующая величину ошибки предсказания на объекте x.*

Для задачи классификации в качестве функции потерь обычно используется индикатор ошибки $L(M) = \operatorname{sign}(M)$.

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

Определение 3. *Эмпирический риск* $Q(X_n, \theta) = \dfrac{1}{n}\sum\limits_{i=1}^{n} L(M_i)$

4. Регуляризация

Проблемы:

- Признаков намного больше, чем объектов
- Мультиколлинеарность признаков:
  Пусть $f(x, \hat{\theta}) = \operatorname{sign}(\theta_1 f_1(x) + \theta_2 f_2(x) - \theta_0)$, $f_2(x) = k f_1(x)$.
  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, у оценки параметров большая дисперсия. Для уменьшения дисперсии можно добавить небольшое смещение, используя регуляризацию.

Задача с регуляризацией:

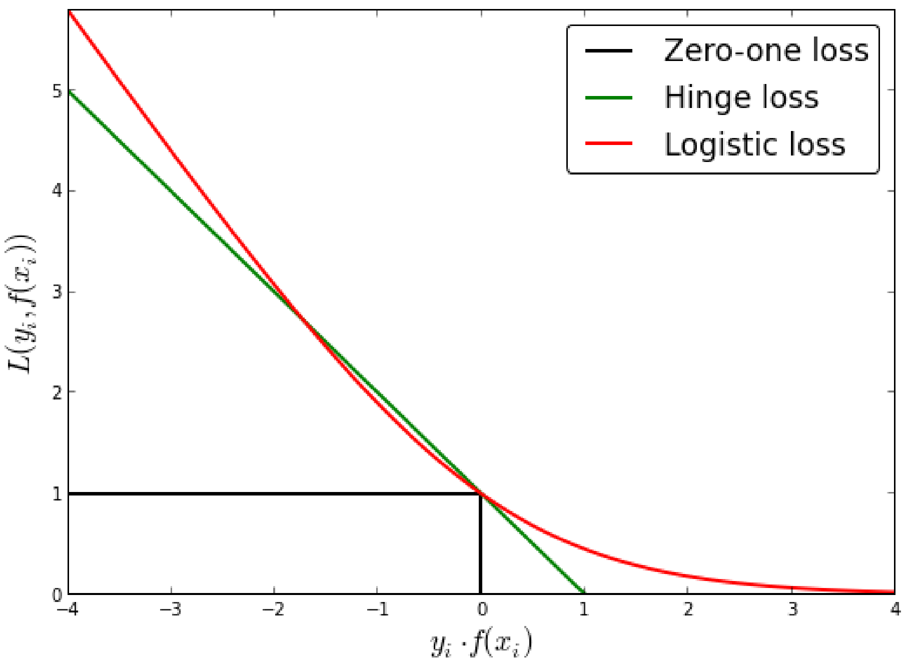
$$\widetilde{Q}(X_n, \theta) = Q(X_n, \theta) + \frac{\tau}{2}\|\theta\|^2 \to \min_{\theta}$$

5. Логистическая регрессия. Подход через минимизацию функции потерь

Линейная модель классификации:

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

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



Задача 1. $\mathsf{Q}(\mathsf{X}_n, \theta) = \sum\limits_{i=1}^{n} \log(1 + \exp(-\mathsf{y}_i\langle\theta, \mathsf{x}_i\rangle)) \to \min\limits_{\theta}$

Методы решения задачи минимизации:

- метод стохастического градиента
- метод Ньютона-Рафсона

6. Логистическая регрессия. Вероятностный подход

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

Свойства $\sigma(z)$:

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

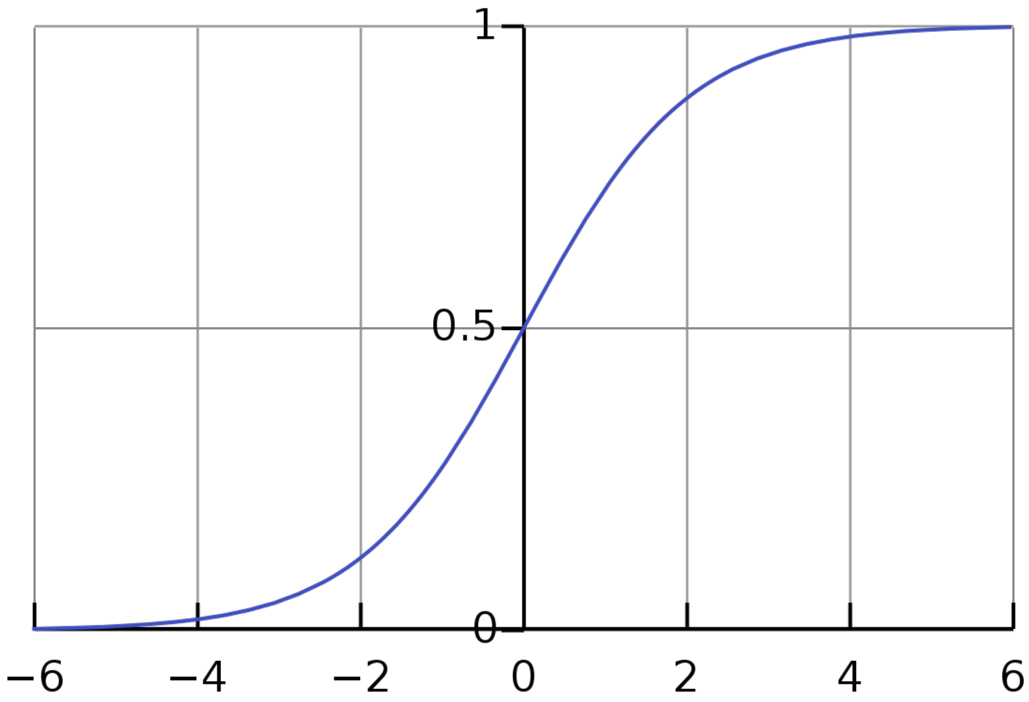


Рис. 1. Сигмоидная функция

Пусть $Y = \{0,1\}$.

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

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

Функция правдоподобия:

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

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

7. Линейная и логистическая регрессия

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

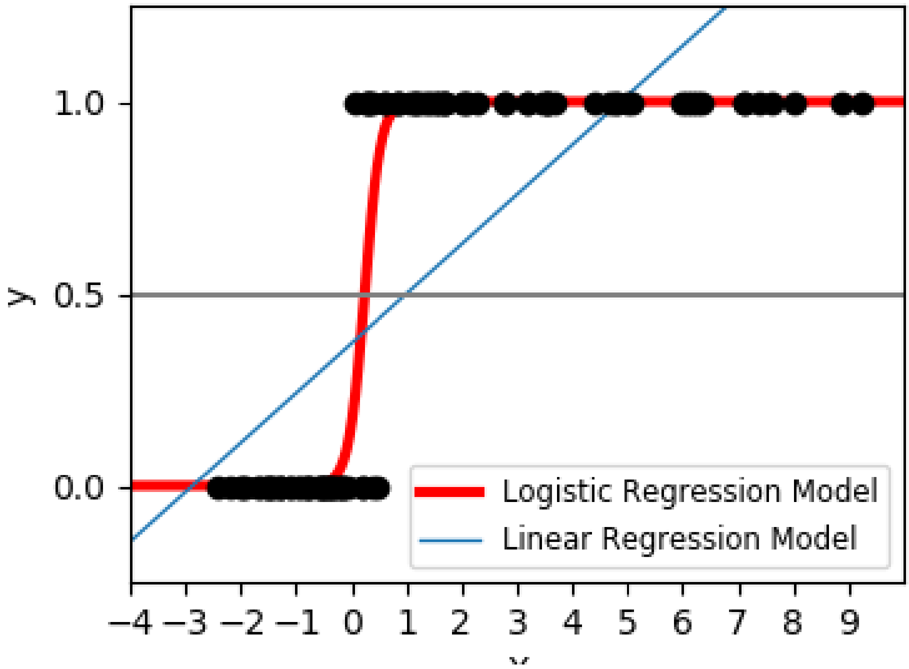


Рис. 2. Линейная и логистическая регрессия

8. Логистическая регрессия. Регуляризация

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

Регуляризация в логистической регрессии:

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

Параметр τ можно подбирать с помощью кросс-валидации.

Методы решения задачи минимизации:

- метод стохастического градиента
- метод Ньютона-Рафсона.

9. 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:

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

10. 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

11. 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