Лекция 4

Логистическая регрессия

1. Введение в логистическую регрессию

Логистическая регрессия – это метод классификации, широко используемый в машинном обучении и статистике для решения задач, где целевая переменная является дискретной, чаще всего бинарной (например, «да» или «нет», «0» или «1»). Логистическая регрессия применяется в различных сферах, таких как медицина, финансы и социология, где важно оценить вероятность возникновения определенного события на основе набора независимых переменных. Несмотря на название, логистическая регрессия относится не к регрессионным, а к классификационным методам.

Логистическая регрессия является популярным методом классификации из-за своей простоты, способности интерпретировать результаты и устойчивости к переобучению на небольших выборках.

2. Основные принципы и формулировка модели

Логистическая регрессия опирается на логистическую функцию (или сигмоидную функцию) для моделирования вероятности принадлежности к одному из классов. Эта функция описывается следующей формулой:

σ(z)=11+e−z\sigma(z) = \frac{1}{1 + e^{-z}}σ(z)=1+e−z1

где zzz — линейная комбинация входных переменных. Сигмоидная функция принимает любые действительные числа в качестве аргументов, но преобразует их в значения на интервале от 0 до 1, что удобно для интерпретации результата как вероятности. Для бинарной классификации целевая переменная yyy принимает значение 1 или 0, и задача заключается в предсказании вероятности P(y=1∣X)P(y=1|X)P(y=1∣X), где XXX — вектор признаков.

Модель логистической регрессии:

Логистическая регрессия моделирует вероятность класса 1 как:

P(y=1∣X)=σ(β0+β1x1+β2x2+...+βnxn)P(y=1|X) = \sigma(\beta_0 + \beta_1x_1 + \beta_2x_2 + ... + \beta_nx_n)P(y=1∣X)=σ(β0+β1x1+β2x2+...+βnxn)

где β0\beta_0β0 — свободный член, а β1,β2,...,βn\beta_1, \beta_2, ..., \beta_nβ1,β2,...,βn — коэффициенты, которые определяют вклад каждой независимой переменной x1,x2,...,xnx_1, x_2, ..., x_nx1,x2,...,xn в предсказание результата.

3. Максимизация правдоподобия

Для оценки коэффициентов логистической регрессии используется метод максимального правдоподобия. Мы выбираем такие значения коэффициентов, которые максимизируют вероятность наблюдаемых значений целевой переменной. Правдоподобие представляется следующим образом:

L(β)=∏i=1mP(yi|Xi;β)yi(1−P(yi|Xi;β))1−yiL(\beta) = \prod_{i=1}^{m} P(y_i | X_i; \beta)^{y_i} (1 - P(y_i | X_i; \beta))^{1 - y_i}L(β)=i=1∏mP(yi|Xi;β)yi(1−P(yi|Xi;β))1−yi

где mmm — количество наблюдений в данных.

Для упрощения расчетов используют логарифм функции правдоподобия:

LogL(β)=∑i=1m[yilog⁡P(yi|Xi;β)+(1−yi)log⁡(1−P(yi|Xi;β))]\text{LogL}(\beta) = \sum_{i=1}^{m} \left[y_i \log P(y_i | X_i; \beta) + (1 - y_i) \log (1 - P(y_i | X_i; \beta)) \right]LogL(β)=i=1∑m[yilogP(yi|Xi;β)+(1−yi)log(1−P(yi|Xi;β))]

Максимизация этой функции позволяет определить оптимальные значения параметров β\betaβ, используя численные методы, такие как метод градиентного спуска или метод Ньютона-Рафсона.

4. Интерпретация коэффициентов логистической регрессии

В линейной регрессии коэффициенты интерпретируются как величина изменения целевой переменной при изменении независимой переменной на единицу. В логистической регрессии интерпретация несколько отличается. Коэффициент βi\beta_iβi в логистической модели указывает на то, как изменение признака xix_ixi на единицу влияет на логарифм шансов:

log⁡(P(y=1|X)1−P(y=1|X))=β0+β1x1+β2x2+...+βnxn\log \left(\frac{P(y=1|X)}{1 - P(y=1|X)} \right) = \beta_0 + \beta_1x_1 + \beta_2x_2 + ... + \beta_nx_nlog(1−P(y=1|X)P(y=1|X))=β0+β1x1+β2x2+...+βnxn

Таким образом, коэффициенты логистической регрессии могут быть интерпретированы как изменение логарифма шансов на единицу изменения независимой переменной. Это позволяет оценивать значимость различных признаков в модели.

5. Оценка качества модели

Оценка качества логистической регрессии требует использования специальных метрик, поскольку модель предсказывает вероятность принадлежности к классу, а не конкретное значение. Основные метрики включают:

- **Матрица ошибок (Confusion Matrix)**: Таблица, в которой показаны правильные и неправильные предсказания для каждого класса.
- **Точность (Accuracy)**: Доля правильных предсказаний среди общего числа предсказаний.

Accuracy=TP+TNTP+TN+FP+FN\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}Accuracy=TP+TN+FP+FNTP+TN

- **Полнота (Recall)**: Доля правильно предсказанных положительных примеров среди всех истинных положительных примеров.

Recall=TPTP+FN\text{Recall} = \frac{TP}{TP + FN}Recall=TP+FNTP

- **Точность (Precision)**: Доля правильных положительных предсказаний среди всех предсказанных положительных примеров.

Precision=TPTP+FP\text{Precision} = \frac{TP}{TP + FP}Precision=TP+FPTP

- **F-мера**: Гармоническое среднее между точностью и полнотой, полезно при неравномерных данных.

F1=2·Precision·RecallPrecision+RecallF_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}F1 =2·Precision+RecallPrecision·Recall

- **Кривая ROC и AUC**: ROC-кривая показывает соотношение между полнотой и ложноположительными срабатываниями при разных порогах, AUC — площадь под ROC-кривой, измеряющая общую производительность модели.

6. Многоклассовая логистическая регрессия

Логистическая регрессия естественным образом расширяется на многоклассовую классификацию. Наиболее распространенные подходы включают:

- **Метод «один против всех» (One-vs-Rest)**: Для каждого класса обучается отдельный классификатор, который предсказывает принадлежность наблюдения к одному конкретному классу.

- **Метод «один против одного» (One-vs-One)**: Создается классификатор для каждой пары классов, а затем большинство голосов используется для определения окончательного класса.

Многоклассовая логистическая регрессия часто используется для задач, где классов несколько, например, для классификации категорий изображений или распознавания рукописных символов.

7. Применение логистической регрессии

Логистическая регрессия находит применение в задачах бинарной и многоклассовой классификации:

- **Медицина**: Прогнозирование вероятности наличия заболевания на основе медицинских данных.
- **Маркетинг**: Оценка вероятности отклика клиента на рекламную кампанию.
- **Финансы**: Прогнозирование вероятности дефолта по кредиту или других рисковых событий.
- **Социология и психология**: Определение вероятности принадлежности к определенным социальным группам или поведенческим категориям.

8. Ограничения логистической регрессии

Логистическая регрессия, несмотря на свою популярность, имеет несколько ограничений:

- **Линейность предположения**: Предполагает линейную зависимость между независимыми переменными и логарифмом шансов, что не всегда соответствует действительности.
- **Неустойчивость к выбросам**: Аномальные значения в данных могут значительно повлиять на результаты.
- **Проблемы мультиколлинеарности**: При наличии высоко коррелированных переменных коэффициенты могут стать нестабильными и сложно интерпретируемыми.

9. Регуляризация в логистической регрессии

Регуляризация является методом борьбы с переобучением, который может быть включен в логистическую регрессию. Два основных типа регуляризации — это L1 (Lasso) и L2 (Ridge):

- **L1-регуляризация (Lasso)**: Добавляет к функции потерь сумму абсолютных значений коэффициентов. Она способствует занулению коэффициентов малозначимых признаков, что также выполняет роль отбора признаков.

Минимизировать −LogL(β)+λ∑|βi|\text{Минимизировать } -\text{LogL}(\beta) + \lambda \sum |\beta_i|Минимизировать −LogL(β)+λ∑|βi|

- **L2-регуляризация (Ridge)**: Добавляет сумму квадратов коэффициентов, что приводит к сглаживанию больших значений коэффициентов, снижая риск переобучения.

Минимизировать −LogL(β)+λ∑βi2\text{Минимизировать } -\text{LogL}(\beta) + \lambda \sum \beta_i^2Минимизировать −LogL(β)+λ∑βi2

10. Заключение

Логистическая регрессия — это мощный и универсальный метод классификации, который, благодаря своей интерпретируемости и простоте, широко используется в различных областях. Хотя логистическая регрессия имеет ограничения и не всегда подходит для сложных нелинейных данных, она продолжает оставаться основным инструментом в арсенале специалистов по анализу данных.