Упражнение 3: Справочник по теории и терминам

Даниил Ковех

2025-10-28

Содержание

1. Как читать этот конспект

Здесь собрано всё базовое, что нужно для задач упражнения 3. Каждый термин объясняю коротко, глаголами и без воды. Формулы даю компактно и помечаю, к какой задаче они относятся.

2. Линейная регрессия и ошибки

2.1. Модель

Что делаем. Описываем зависимость численного ответа y от матрицы признаков X:

$$y = X\beta + \varepsilon.$$

- β — вектор коэффициентов.

- ε — шум: случайные ошибки прогноза.
- X — фиксированный или случайный дизайн (в задачах 1–3 берём фиксированный).

2.2. Оценка методом наименьших квадратов (OLS)

Задача. Найти $\hat{\beta}$, минимизируя сумму квадратов остатков:

$$\hat{\beta} = (X^\top X)^{-1} X^\top y.$$

Зачем. Это базовая точка отсчёта для всех последующих отклонений, регуляризаций и оценок ошибок.

2.3. Обучающая ошибка (training error)

Определяем. Средний квадрат остатка на тех данных, где мы подгоняли модель:

$$\text{err} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2.$$

Особенность. Эта ошибка оптимистична. Мы уже видели эти наблюдения и подстроили под них коэффициенты.

2.4. Внутривыборочная ошибка (in-sample error)

Определяем. Берём те же точки x_i, но новое случайное наблюдение Y_i' и считаем ожидаемый квадрат ошибки:

$$\text{Err}_{\text{in}} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\big[(Y_i' - \hat{f}(x_i))^2\big].$$

Зачем. Хотим понять, как модель поведёт себя на новых откликах при той же сетке признаков.

2.5. Оптимизм (optimism)

Считаем. Разница между ожидаемой внутривыборочной и обучающей ошибкой:

$$O = \text{Err}_{\text{in}} - \mathbb{E}[\text{err}] = \frac{2}{N}\sum_{i=1}^{N}\text{Cov}(\hat{y}_i, y_i).$$

Интерпретируем. Если модель сильно реагирует на шум (большая ковариация), тренировка кажется успешной, но прогнозы будут хуже.

3. Ожидания и ковариации

3.1. Математическое ожидание

Определяем. Среднее значение случайной величины. В линейной модели шум имеет нулевое ожидание: $\mathbb{E}[\varepsilon] = 0$.

3.2. Ковариация

Определяем. Мера совместных колебаний двух случайных величин:

$$\mathrm{Cov}(A, B) = \mathbb{E}[AB] - \mathbb{E}[A]\mathbb{E}[B].$$

Применяем. Формула оптимизма сводится к ковариации между предсказаниями и наблюдениями.

4. Маргинальное правдоподобие и Байес

4.1. Правдоподобие

Определяем. Вероятность данных при фиксированном параметре. Для Бернулли:

$$p(y \mid \theta) = \theta^s (1-\theta)^{N-s}, \quad s = \sum y_i.$$

4.2. Приор Beta

Записываем. Бета-распределение:

$$\mathrm{Beta}(\alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

Используем. Описывает предварительную веру о θ; легко комбинируется с Бернулли.

4.3. Маргинальное правдоподобие

Интегрируем. Убираем параметр, интегрируя по всему диапазону:

$$p(y \mid M) = \int_0^1 p(y \mid \theta) p(\theta)\, d\theta.$$

Формула. Для Beta–Bernoulli:

$$p(y \mid M) = \frac{\mathrm{B}(s+\alpha, N-s+\beta)}{\mathrm{B}(\alpha, \beta)}.$$

4.4. Аппроксимация Лапласа

Идея. Разложить логарифм правдоподобия вокруг максимума и заменить интеграл гауссовым приближением:

$$p(y \mid M) \approx \exp(\ell(\widehat{\theta})) \sqrt{\frac{2\pi}{J(\widehat{\theta})}}.$$

где J — наблюдаемая информация.

Когда. Подходит при большом N и хорошо концентрированном правдоподобии.

4.5. Критерий BIC

Определяем. Приближение к $-2\log p(y \mid M)$:

$$\text{BIC} = -2\ell(\hat{\theta}) + k\log N,$$

где k — число параметров.

Назначение. Быстро сравнивает модели: меньше — лучше.

5. Регуляризация и отбор признаков

5.1. Ridge регрессия

Формула. Добавляем L_2-штраф:

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \frac{1}{N}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \right\}.$$

Эффект. Сжимаем коэффициенты, но не обнуляем. Устраняем мультиколлинеарность.

5.2. Лассо

Формула. Добавляем L_1-штраф:

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \frac{1}{2N}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \right\}.$$

Эффект. Обнуляем часть коэффициентов. Выполняем отбор признаков и интерпретируем результат.

5.3. Параметр λ

Что делает. Управляет силой штрафа. Малое λ — модель похожа на OLS, большое — жёстко ограничивает коэффициенты.

Как выбирать. Кросс-валидация (см. задачу 6) или отложенная выборка (см. задачу 7).

5.4. True Positive Rate (TPR)

Определяем. Доля правильно выбранных значимых признаков:

$$\text{TPR} = \frac{\#\{\text{верно включённых важных}\}}{\#\{\text{всех важных}\}}.$$

5.5. False Positive Rate (FPR)

Определяем. Доля ошибочно включённых шумовых признаков:

$$\text{FPR} = \frac{\#\{\text{включили шум}\}}{\#\{\text{всего шумовых}\}}.$$

6. Кросс-валидация

6.1. LOOCV (Leave-One-Out)

Алгоритм. Для каждого наблюдения:

1. Удаляем i-ю точку.
2. Обучаем модель на оставшихся $N - 1$.
3. Предсказываем y_i.

Формула для линейной модели.

$$\mathrm{CV}_{\mathrm{LOO}} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{y_i - \hat{y}_i}{1 - h_{ii}} \right)^2,$$

где h_{ii} — диагональные элементы матрицы шляп $H = X(X^\top X)^{-1} X^\top$.

6.2. k-fold CV

Алгоритм.

1. Делим выборку на k равных блоков.
2. Обучаем модель на $k - 1$ блоках.
3. Предсказываем на отложенном блоке.
4. Повторяем для всех блоков и усредняем ошибку.

Плюсы. Быстрее, чем LOOCV, даёт меньше дисперсию, если повторить с разными разбивками.

7. Полиномы и контрасты

7.1. Ортогональные полиномы

Что делаем. Преобразуем числовой признак (возраст) в набор ортогональных компонент: `poly(age, 4)`.
Зачем. Сохраняем информацию о степени, но избегаем сильной корреляции между столбцами.

7.2. Difference Contrasts

Что делаем. Для упорядоченных категорий (образование) используем `contr.sdif` из `MASS`.
Поясняем. Каждая колонка кодирует разницу между соседними уровнями. Это удобнее интерпретировать, чем “пустые” дамми.

7.3. Reference Level

Определяем. Категория, с которой сравниваются остальные уровни фактора.
Настройка. Для большинства признаков ставим модальный уровень (самый частый), чтобы оценки интерпретировались как отклонение от типичного случая.

8. Среднеквадратичная ошибка (MSE)

Определяем.

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

В задачах. MSE используется как ключевой критерий при выборе λ и сравнении моделей на тесте.

9. Как использовать справочник

- Проверяйте определения перед решением конкретной задачи.
- Сравнивайте формулы с практическими вычислениями в соответствующих Rmd-файлах.
- Если не хватает интуиции, перечитайте раздел “Жизненный пример” в файле по соответствующей задаче.

Этот конспект — быстрый способ вспомнить, что означает каждый термин, и как он появляется в формулах упражнений.