Байесовские Методы. Лекция 3

БИВТ-23-9/10-ИСАД

23 сентября 2025 г.

Классическая линейная регрессия

Модель

$$y = X\theta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

- X — матрица признаков размера $n \times d$.

Классическая линейная регрессия

Модель

$$y = X\theta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

- X — матрица признаков размера $n \times d$.
- θ — вектор коэффициентов (неизвестные параметры).

Классическая линейная регрессия

Модель

$$y = X\theta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

- X — матрица признаков размера $n \times d$.
- θ — вектор коэффициентов (неизвестные параметры).
- y — вектор наблюдений.

Классическая линейная регрессия

Модель

$$y = X\theta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

- X — матрица признаков размера $n \times d$.
- θ — вектор коэффициентов (неизвестные параметры).
- y — вектор наблюдений.
- Ошибки ε независимы, одинаково распределены, дисперсия σ^2.

Цель

Найти такие параметры θ, которые лучше всего объясняют данные.

Оценка параметров (MLE)

- Правдоподобие:

$$P(y \mid X, \theta) = \prod_{i=1}^{n} \mathcal{N}(y_i \mid x_i^\top \theta, \sigma^2).$$

Оценка параметров (MLE)

- Правдоподобие:

$$P(y \mid X, \theta) = \prod_{i=1}^{n} \mathcal{N}(y_i \mid x_i^\top \theta, \sigma^2).$$

- Лог-правдоподобие:

$$\ell(\theta) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|y - X\theta\|^2.$$

Оценка параметров (MLE)

- Правдоподобие:

$$P(y \mid X, \theta) = \prod_{i=1}^{n} \mathcal{N}(y_i \mid x_i^\top \theta, \sigma^2).$$

- Лог-правдоподобие:

$$\ell(\theta) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|y - X\theta\|^2.$$

- Максимизация эквивалентна задаче МНК:

$$\hat{\theta}_{\mathrm{MLE}} = \arg\min_{\theta} \|y - X\theta\|^2.$$

Вывод решения для линейной регрессии

Оптимизационная задача (MLE)

$$\hat{\theta}_{\mathsf{MLE}} = \arg\min_{\theta} \|y - X\theta\|^2$$

- Запишем функцию ошибки (сумму квадратов):

$$J(\theta) = (y - X\theta)^\top (y - X\theta).$$

Вывод решения для линейной регрессии

Оптимизационная задача (MLE)

$$\hat{\theta}_{\text{MLE}} = \arg\min_{\theta} \|y - X\theta\|^2$$

- Запишем функцию ошибки (сумму квадратов):

$$J(\theta) = (y - X\theta)^\top (y - X\theta).$$

- Берём градиент по θ:

$$\nabla_\theta J(\theta) = -2X^\top (y - X\theta).$$

Вывод решения для линейной регрессии

Оптимизационная задача (MLE)

$$\hat{\theta}_{\text{MLE}} = \arg\min_{\theta} \|y - X\theta\|^2$$

- Запишем функцию ошибки (сумму квадратов):

$$J(\theta) = (y - X\theta)^\top (y - X\theta).$$

- Берём градиент по θ:

$$\nabla_\theta J(\theta) = -2X^\top (y - X\theta).$$

- Приравниваем к нулю:

$$X^\top X \hat{\theta} = X^\top y.$$

Вывод решения для линейной регрессии

Оптимизационная задача (MLE)

$$\hat{\theta}_{\mathsf{MLE}} = \arg\min_{\theta} \|y - X\theta\|^2$$

- Запишем функцию ошибки (сумму квадратов):

$$J(\theta) = (y - X\theta)^\top (y - X\theta).$$

- Берём градиент по θ:

$$\nabla_\theta J(\theta) = -2X^\top (y - X\theta).$$

- Приравниваем к нулю:

$$X^\top X \hat{\theta} = X^\top y.$$

- Если $X^\top X$ обратима:

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

Нормальные уравнения и псевдообратная

Нормальные уравнения

$$X^\top X\hat{\theta} = X^\top y$$

- Если $X^\top X$ обратима:

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

Нормальные уравнения и псевдообратная

Нормальные уравнения

$$X^\top X \hat{\theta} = X^\top y$$

- Если $X^\top X$ обратима:
$$\hat{\theta} = (X^\top X)^{-1} X^\top y.$$
- Если $X^\top X$ вырожденная (нет обратной), решение выражается через **псевдообратную Мура–Пенроуза**:
$$\hat{\theta} = X^+ y,$$
где
$$X^+ = \begin{cases} (X^\top X)^{-1} X^\top, & n \geq d,\ \text{ранг}(X) = d, \\ X^\top (X X^\top)^{-1}, & n < d,\ \text{ранг}(X) = n. \end{cases}$$

Нормальные уравнения

$$X^\top X\hat{\theta} = X^\top y$$

- Если $X^\top X$ обратима:
$$\hat{\theta} = (X^\top X)^{-1}X^\top y.$$
- Если $X^\top X$ вырожденная (нет обратной), решение выражается через **псевдообратную Мура–Пенроуза**:
$$\hat{\theta} = X^+ y,$$
где
$$X^+ = \begin{cases} (X^\top X)^{-1}X^\top, & n \geq d,\ \text{ранг}(X) = d, \\ X^\top (XX^\top)^{-1}, & n < d,\ \text{ранг}(X) = n. \end{cases}$$
- Это решение минимизирует $\|y - X\theta\|^2$ и среди всех возможных выбирает θ с минимальной нормой.

Нормальные уравнения и псевдообратная

Нормальные уравнения

$$X^\top X\hat{\theta} = X^\top y$$

- Если $X^\top X$ обратима:
$$\hat{\theta} = (X^\top X)^{-1} X^\top y.$$
- Если $X^\top X$ вырожденная (нет обратной), решение выражается через **псевдообратную Мура–Пенроуза**:
$$\hat{\theta} = X^+ y,$$
где
$$X^+ = \begin{cases} (X^\top X)^{-1} X^\top, & n \geq d,\ \text{ранг}(X) = d, \\ X^\top (XX^\top)^{-1}, & n < d,\ \text{ранг}(X) = n. \end{cases}$$
- Это решение минимизирует $\|y - X\theta\|^2$ и среди всех возможных выбирает θ с минимальной нормой.

Замечание

Использование псевдообратной особенно важно при $n < d$ (мало данных, много признаков).

Ограничения классической регрессии (MLE)

- **Нет априорной информации:** модель не учитывает знаний о параметрах до наблюдений.

Ограничения классической регрессии (MLE)

- **Нет априорной информации:** модель не учитывает знаний о параметрах до наблюдений.
- **Точка вместо распределения:** $\hat{\theta}$ — одно число, без оценки неопределённости.

Ограничения классической регрессии (MLE)

- **Нет априорной информации:** модель не учитывает знаний о параметрах до наблюдений.
- **Точка вместо распределения:** $\hat{\theta}$ — одно число, без оценки неопределённости.
- **Переобучение:** при $n < d$ или при сильной корреляции признаков решение становится нестабильным.

Ограничения классической регрессии (MLE)

- **Нет априорной информации:** модель не учитывает знаний о параметрах до наблюдений.
- **Точка вместо распределения:** $\hat{\theta}$ — одно число, без оценки неопределённости.
- **Переобучение:** при $n < d$ или при сильной корреляции признаков решение становится нестабильным.
- **Нет доверительных интервалов:** предсказания не содержат информацию о «надежности».

Почему байесовская регрессия?

- Добавляем **априор** на параметры θ:

$$\theta \sim \mathcal{N}(0, \tau^2 I).$$

Почему байесовская регрессия?

- Добавляем **априор** на параметры θ:

$$\theta \sim \mathcal{N}(0, \tau^2 I).$$

- Получаем **постериорное распределение**, а не точку:

$$P(\theta \mid D) \propto P(y \mid X, \theta)\, P(\theta).$$

Почему байесовская регрессия?

- Добавляем **априор** на параметры θ:

$$\theta \sim \mathcal{N}(0, \tau^2 I).$$

- Получаем **постериорное распределение**, а не точку:

$$P(\theta \mid D) \propto P(y \mid X, \theta) \, P(\theta).$$

- Теперь можно делать **байесовское предсказание**:

$$P(y^* \mid x^*, D) = \int P(y^* \mid x^*, \theta) \, P(\theta \mid D) \, d\theta.$$

Байесовская линейная регрессия: модель

Наблюдения

$$y = X\theta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I).$$

Априор на параметры

$$\theta \sim \mathcal{N}(0, \tau^2 I).$$

- Правдоподобие:

$$P(y \mid X, \theta) = \mathcal{N}(y \mid X\theta, \sigma^2 I).$$

Байесовская линейная регрессия: модель

Наблюдения

$$y = X\theta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I).$$

Априор на параметры

$$\theta \sim \mathcal{N}(0, \tau^2 I).$$

- Правдоподобие:
$$P(y \mid X, \theta) = \mathcal{N}(y \mid X\theta, \sigma^2 I).$$
- Априор: гауссовский, центрирован в нуле.

Байесовская линейная регрессия: модель

Наблюдения

$$y = X\theta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I).$$

Априор на параметры

$$\theta \sim \mathcal{N}(0, \tau^2 I).$$

- Правдоподобие:
$$P(y \mid X, \theta) = \mathcal{N}(y \mid X\theta, \sigma^2 I).$$
- Априор: гауссовский, центрирован в нуле.
- Постериор: по теореме Байеса

$$P(\theta \mid D) \propto P(y \mid X, \theta)\, P(\theta).$$

Вывод постериора для параметров

Теорема Байеса

$$P(\theta \mid D) \propto P(y \mid X, \theta)\, P(\theta).$$

- Правдоподобие:

$$P(y \mid X, \theta) \propto \exp\left(-\frac{1}{2\sigma^2}\|y - X\theta\|^2\right).$$

Вывод постериора для параметров

Теорема Байеса

$$P(\theta \mid D) \propto P(y \mid X, \theta)\, P(\theta).$$

- Правдоподобие:

$$P(y \mid X, \theta) \propto \exp\left(-\frac{1}{2\sigma^2} \|y - X\theta\|^2 \right).$$

- Априор:

$$P(\theta) \propto \exp\left(-\frac{1}{2\tau^2} \|\theta\|^2 \right).$$

Вывод постериора для параметров

Теорема Байеса

$$P(\theta \mid D) \propto P(y \mid X, \theta)\, P(\theta).$$

- Правдоподобие:
$$P(y \mid X, \theta) \propto \exp\Big(-\tfrac{1}{2\sigma^2} \|y - X\theta\|^2 \Big).$$
- Априор:
$$P(\theta) \propto \exp\Big(-\tfrac{1}{2\tau^2} \|\theta\|^2 \Big).$$
- Постериор:
$$P(\theta \mid D) \propto \exp\Big(-\tfrac{1}{2\sigma^2} \|y - X\theta\|^2 - \tfrac{1}{2\tau^2} \|\theta\|^2 \Big).$$

Вывод постериора для параметров

Теорема Байеса

$$P(\theta \mid D) \propto P(y \mid X, \theta)\, P(\theta).$$

- Правдоподобие:
$$P(y \mid X, \theta) \propto \exp\Big(-\tfrac{1}{2\sigma^2} \|y - X\theta\|^2 \Big).$$
- Априор:
$$P(\theta) \propto \exp\Big(-\tfrac{1}{2\tau^2} \|\theta\|^2 \Big).$$
- Постериор:
$$P(\theta \mid D) \propto \exp\Big(-\tfrac{1}{2\sigma^2} \|y - X\theta\|^2 - \tfrac{1}{2\tau^2} \|\theta\|^2 \Big).$$
- Это экспонента от квадратичной формы по θ → многомерное нормальное распределение.

Вывод постериора для параметров

Теорема Байеса

$$P(\theta \mid D) \propto P(y \mid X, \theta)\, P(\theta).$$

- Правдоподобие:
$$P(y \mid X, \theta) \propto \exp\left(-\frac{1}{2\sigma^2} \|y - X\theta\|^2 \right).$$
- Априор:
$$P(\theta) \propto \exp\left(-\frac{1}{2\tau^2} \|\theta\|^2 \right).$$
- Постериор:
$$P(\theta \mid D) \propto \exp\left(-\frac{1}{2\sigma^2} \|y - X\theta\|^2 - \frac{1}{2\tau^2} \|\theta\|^2 \right).$$
- Это экспонента от квадратичной формы по θ → многомерное нормальное распределение.

Результат

$$P(\theta \mid D) = \mathcal{N}(\mu_N, \Sigma_N),$$

где

$$\Sigma_N = \left(\frac{1}{\sigma^2} X^\top X + \frac{1}{\tau^2} I \right)^{-1}, \quad \mu_N = \frac{1}{\sigma^2} \Sigma_N X^\top y.$$

Интуиция: что означают формулы?

- Σ_N — ковариация постериора.

Интуиция: что означают формулы?

- Σ_N — ковариация постериора.
  - Чем больше данных n, тем больше $X^\top X$, тем меньше дисперсия → растёт уверенность.

Интуиция: что означают формулы?

- Σ_N — ковариация постериора.
    - Чем больше данных n, тем больше $X^\top X$, тем меньше дисперсия → растёт уверенность.
    - Малое τ^2 (сильный априор) уменьшает вариацию → сжатие к нулю.

Интуиция: что означают формулы?

- Σ_N — ковариация постериора.
  - Чем больше данных n, тем больше $X^\top X$, тем меньше дисперсия → растёт уверенность.
  - Малое τ^2 (сильный априор) уменьшает вариацию → сжатие к нулю.
- μ_N — центр постериора.

Интуиция: что означают формулы?

Постериор

$$P(\theta \mid D) = \mathcal{N}(\mu_N, \Sigma_N),$$

где

$$\Sigma_N = \left(\frac{1}{\sigma^2}X^\top X + \frac{1}{\tau^2}I\right)^{-1}, \quad \mu_N = \frac{1}{\sigma^2}\Sigma_N X^\top y.$$

- Σ_N — ковариация постериора.
  - Чем больше данных n, тем больше $X^\top X$, тем меньше дисперсия → растёт уверенность.
  - Малое τ^2 (сильный априор) уменьшает вариацию → сжатие к нулю.
- μ_N — центр постериора.
  - Баланс между MLE и априором.

Интуиция: что означают формулы?

Постериор

$$P(\theta \mid D) = \mathcal{N}(\mu_N, \Sigma_N),$$

где

$$\Sigma_N = \left(\frac{1}{\sigma^2} X^\top X + \frac{1}{\tau^2} I\right)^{-1}, \quad \mu_N = \frac{1}{\sigma^2} \Sigma_N X^\top y.$$

- Σ_N — ковариация постериора.
  - Чем больше данных n, тем больше $X^\top X$, тем меньше дисперсия → растёт уверенность.
  - Малое τ^2 (сильный априор) уменьшает вариацию → сжатие к нулю.
- μ_N — центр постериора.
  - Баланс между MLE и априором.
  - При $\tau^2 \to \infty$ (слабый априор): $\mu_N \to \hat{\theta}_{\mathrm{MLE}}$.

Интуиция: что означают формулы?

Постериор

$$P(\theta \mid D) = \mathcal{N}(\mu_N, \Sigma_N),$$

где

$$\Sigma_N = \left(\frac{1}{\sigma^2}X^\top X + \frac{1}{\tau^2}I\right)^{-1}, \quad \mu_N = \frac{1}{\sigma^2}\Sigma_N X^\top y.$$

- Σ_N — ковариация постериора.
  - Чем больше данных n, тем больше $X^\top X$, тем меньше дисперсия → растёт уверенность.
  - Малое τ^2 (сильный априор) уменьшает вариацию → сжатие к нулю.
- μ_N — центр постериора.
  - Баланс между MLE и априором.
  - При $\tau^2 \to \infty$ (слабый априор): $\mu_N \to \hat{\theta}_{\text{MLE}}$.
  - При малом τ^2: μ_N тяготеет к нулю.

Интуиция: что означают формулы?

Постериор

$$P(\theta \mid D) = \mathcal{N}(\mu_N, \Sigma_N),$$

где

$$\Sigma_N = \left(\frac{1}{\sigma^2} X^\top X + \frac{1}{\tau^2} I\right)^{-1}, \quad \mu_N = \frac{1}{\sigma^2} \Sigma_N X^\top y.$$

- Σ_N — ковариация постериора.
  - Чем больше данных n, тем больше $X^\top X$, тем меньше дисперсия → растёт уверенность.
  - Малое τ^2 (сильный априор) уменьшает вариацию → сжатие к нулю.
- μ_N — центр постериора.
  - Баланс между MLE и априором.
  - При $\tau^2 \to \infty$ (слабый априор): $\mu_N \to \hat{\theta}_{\text{MLE}}$.
  - При малом τ^2: μ_N тяготеет к нулю.
- В целом:

Интуиция: что означают формулы?

Постериор

$$P(\theta \mid D) = \mathcal{N}(\mu_N, \Sigma_N),$$

где

$$\Sigma_N = \left(\frac{1}{\sigma^2} X^\top X + \frac{1}{\tau^2} I\right)^{-1}, \quad \mu_N = \frac{1}{\sigma^2} \Sigma_N X^\top y.$$

- Σ_N — ковариация постериора.
  - Чем больше данных n, тем больше $X^\top X$, тем меньше дисперсия → растёт уверенность.
  - Малое τ^2 (сильный априор) уменьшает вариацию → сжатие к нулю.
- μ_N — центр постериора.
  - Баланс между MLE и априором.
  - При $\tau^2 \to \infty$ (слабый априор): $\mu_N \to \hat{\theta}_{\text{MLE}}$.
  - При малом τ^2: μ_N тяготеет к нулю.
- В целом:
  - Байесовская регрессия = MLE + априор -> неопределённость.

Интуиция: что означают формулы?

Постериор

$$P(\theta \mid D) = \mathcal{N}(\mu_N, \Sigma_N),$$

где

$$\Sigma_N = \left(\frac{1}{\sigma^2} X^\top X + \frac{1}{\tau^2} I\right)^{-1}, \quad \mu_N = \frac{1}{\sigma^2} \Sigma_N X^\top y.$$

- Σ_N — ковариация постериора.
  - Чем больше данных n, тем больше $X^\top X$, тем меньше дисперсия → растёт уверенность.
  - Малое τ^2 (сильный априор) уменьшает вариацию → сжатие к нулю.
- μ_N — центр постериора.
  - Баланс между MLE и априором.
  - При $\tau^2 \to \infty$ (слабый априор): $\mu_N \to \hat{\theta}_{\text{MLE}}$.
  - При малом τ^2: μ_N тяготеет к нулю.
- В целом:
  - Байесовская регрессия = MLE + априор -> неопределённость.
  - Мы получаем распределение параметров, а не одно число.

Байесовское предсказание

Определение

Для нового объекта x^*:

$$P(y^* \mid x^*, D) = \int P(y^* \mid x^*, \theta, \sigma^2)\, P(\theta \mid D)\, d\theta.$$

- Внутренний интеграл усредняет предсказания по всем возможным θ.
- Благодаря сопряжённости (нормальное $\times$ нормальное) интеграл считается аналитически.

Байесовское предсказание

Определение

Для нового объекта x^*:

$$P(y^* \mid x^*, D) = \int P(y^* \mid x^*, \theta, \sigma^2)\, P(\theta \mid D)\, d\theta.$$

- Внутренний интеграл усредняет предсказания по всем возможным θ.
- Благодаря сопряжённости (нормальное $\times$ нормальное) интеграл считается аналитически.

Байесовское предсказание

Определение

Для нового объекта x^*:

$$P(y^* \mid x^*, D) = \int P(y^* \mid x^*, \theta, \sigma^2)\, P(\theta \mid D)\, d\theta.$$

- Внутренний интеграл усредняет предсказания по всем возможным θ.
- Благодаря сопряжённости (нормальное $\times$ нормальное) интеграл считается аналитически.

Результат

$$P(y^* \mid x^*, D) = \mathcal{N}\Big(x^{*\top}\mu_N,\; x^{*\top}\Sigma_N x^* + \sigma^2\Big),$$

где μ_N, Σ_N — параметры постериора для θ.

MAP-оценка $\Leftrightarrow$ Ridge: вывод

Модель и априор

Линейная регрессия: $y = X\theta + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$
Гауссовский априор на коэффициенты: $\theta \sim \mathcal{N}(0, \tau^2 I)$

- Постериор по теореме Байеса:

$$P(\theta \mid D) \propto P(y \mid X, \theta)\, P(\theta) \propto \exp\Big(- \tfrac{1}{2\sigma^2} \|y - X\theta\|^2 - \tfrac{1}{2\tau^2} \|\theta\|^2 \Big).$$

MAP-оценка $\Leftrightarrow$ Ridge: вывод

Модель и априор

Линейная регрессия: $y = X\theta + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$
Гауссовский априор на коэффициенты: $\theta \sim \mathcal{N}(0, \tau^2 I)$

- Постериор по теореме Байеса:

$$P(\theta \mid D) \propto P(y \mid X, \theta)\, P(\theta) \propto \exp\Big(-\tfrac{1}{2\sigma^2} \|y - X\theta\|^2 - \tfrac{1}{2\tau^2} \|\theta\|^2 \Big).$$

- **MAP** = максимум постериора = минимум отрицательного лог-постериора:

$$\hat{\theta}_{\mathrm{MAP}} = \arg\min_{\theta} \left[\tfrac{1}{2\sigma^2} \|y - X\theta\|^2 + \tfrac{1}{2\tau^2} \|\theta\|^2 \right].$$

MAP-оценка $\Leftrightarrow$ Ridge: вывод

Модель и априор

Линейная регрессия: $y = X\theta + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$
Гауссовский априор на коэффициенты: $\theta \sim \mathcal{N}(0, \tau^2 I)$

- Постериор по теореме Байеса:

$$P(\theta \mid D) \propto P(y \mid X, \theta)\, P(\theta) \propto \exp\Big(- \tfrac{1}{2\sigma^2} \|y - X\theta\|^2 - \tfrac{1}{2\tau^2} \|\theta\|^2 \Big).$$

- **MAP** = максимум постериора = минимум отрицательного лог-постериора:

$$\hat{\theta}_{\mathsf{MAP}} = \arg\min_{\theta} \Big[\tfrac{1}{2\sigma^2} \|y - X\theta\|^2 + \tfrac{1}{2\tau^2} \|\theta\|^2 \Big].$$

- Умножим на $2\sigma^2$ (не меняет минимума) и положим $\lambda = \sigma^2/\tau^2$:

$$\hat{\theta}_{\mathsf{MAP}} = \arg\min_{\theta}\ \|y - X\theta\|^2 + \lambda \|\theta\|^2.$$

MAP-оценка $\Leftrightarrow$ Ridge: вывод

Модель и априор

Линейная регрессия: $y = X\theta + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$
Гауссовский априор на коэффициенты: $\theta \sim \mathcal{N}(0, \tau^2 I)$

- Постериор по теореме Байеса:

$$P(\theta \mid D) \propto P(y \mid X, \theta)\, P(\theta) \propto \exp\Big(- \tfrac{1}{2\sigma^2} \|y - X\theta\|^2 - \tfrac{1}{2\tau^2} \|\theta\|^2 \Big).$$

- **MAP** = максимум постериора = минимум отрицательного лог-постериора:

$$\hat{\theta}_{\mathrm{MAP}} = \arg\min_{\theta} \Big[\tfrac{1}{2\sigma^2} \|y - X\theta\|^2 + \tfrac{1}{2\tau^2} \|\theta\|^2 \Big].$$

- Умножим на $2\sigma^2$ (не меняет минимума) и положим $\lambda = \sigma^2/\tau^2$:

$$\hat{\theta}_{\mathrm{MAP}} = \arg\min_{\theta}\ \|y - X\theta\|^2 + \lambda \|\theta\|^2.$$

- Это в точности задача **Ridge regression** (Тихоновская регуляризация, L_2):

$$\boxed{\min_{\theta}\ \|y - X\theta\|^2 + \lambda \|\theta\|^2}.$$

Выбор λ через evidence

Напомним:

MAP для гауссовского априора эквивалентен Ridge:

$$\hat{\theta}_{\mathsf{MAP}} = (X^\top X + \lambda I)^{-1} X^\top y, \quad \lambda = \frac{\sigma^2}{\tau^2}.$$

- Вопрос: как выбрать λ?

Выбор λ через evidence

Напомним:

MAP для гауссовского априора эквивалентен Ridge:

$$\hat{\theta}_{\text{MAP}} = (X^\top X + \lambda I)^{-1} X^\top y, \quad \lambda = \frac{\sigma^2}{\tau^2}.$$

- Вопрос: как выбрать λ?
- Частотный подход: кросс-валидация.

Выбор λ через evidence

Напомним:

MAP для гауссовского априора эквивалентен Ridge:

$$\hat{\theta}_{\mathrm{MAP}} = (X^\top X + \lambda I)^{-1} X^\top y, \quad \lambda = \frac{\sigma^2}{\tau^2}.$$

- Вопрос: как выбрать λ?
- Частотный подход: кросс-валидация.
- Байесовский подход: **максимизация evidence**.

$$P(y \mid X, \lambda, \sigma^2) = \int P(y \mid X, \theta, \sigma^2)\, P(\theta \mid \lambda)\, d\theta.$$

Evidence для линейной регрессии

Сопряжённость (нормальное $\times$ нормальное)

Интеграл берётся в замкнутой форме:

$$P(y \mid X, \lambda, \sigma^2) = \mathcal{N}\big(y \mid 0,\ \sigma^2 I + \tau^2 X X^\top\big).$$

- В терминах $\lambda = \sigma^2/\tau^2$:

$$\log P(y \mid X, \lambda, \sigma^2) = -\tfrac{1}{2}\Big(n \log(2\pi) + \log\det C + y^\top C^{-1} y\Big),$$

где $C = \sigma^2 I + \tau^2 X X^\top$.

Evidence для линейной регрессии

Сопряжённость (нормальное $\times$ нормальное)

Интеграл берётся в замкнутой форме:

$$P(y \mid X, \lambda, \sigma^2) = \mathcal{N}\big(y \mid 0,\ \sigma^2 I + \tau^2 X X^\top\big).$$

- В терминах $\lambda = \sigma^2/\tau^2$:

$$\log P(y \mid X, \lambda, \sigma^2) = -\frac{1}{2}\Big(n \log(2\pi) + \log\det C + y^\top C^{-1} y\Big),$$

  где $C = \sigma^2 I + \tau^2 X X^\top$.
- Задача: выбрать $\hat{\lambda} = \arg\max_\lambda \log P(y \mid X, \lambda, \sigma^2)$.

Интуиция выбора λ через evidence

- Малое λ (слабый априор, τ^2 большое) $\Rightarrow$ модель гибкая, риск переобучения.

Интуиция выбора λ через evidence

- Малое λ (слабый априор, τ^2 большое) $\Rightarrow$ модель гибкая, риск переобучения.
- Большое λ (сильный априор, τ^2 маленькое) $\Rightarrow$ коэффициенты сжаты к нулю, риск недообучения.

Интуиция выбора λ через evidence

- Малое λ (слабый априор, τ^2 большое) $\Rightarrow$ модель гибкая, риск переобучения.
- Большое λ (сильный априор, τ^2 маленькое) $\Rightarrow$ коэффициенты сжаты к нулю, риск недообучения.
- Evidence $P(y \mid X, \lambda)$ балансирует эти два эффекта:

Интуиция выбора λ через evidence

- Малое λ (слабый априор, τ^2 большое) $\Rightarrow$ модель гибкая, риск переобучения.
- Большое λ (сильный априор, τ^2 маленькое) $\Rightarrow$ коэффициенты сжаты к нулю, риск недообучения.
- Evidence $P(y \mid X, \lambda)$ балансирует эти два эффекта:
  - штрафует за избыточную сложность (бритва Оккама),

Интуиция выбора λ через evidence

- Малое λ (слабый априор, τ^2 большое) $\Rightarrow$ модель гибкая, риск переобучения.
- Большое λ (сильный априор, τ^2 маленькое) $\Rightarrow$ коэффициенты сжаты к нулю, риск недообучения.
- Evidence $P(y \mid X, \lambda)$ балансирует эти два эффекта:
  - штрафует за избыточную сложность (бритва Оккама),
  - поощряет модели, которые хорошо объясняют данные.

Интуиция выбора λ через evidence

- Малое λ (слабый априор, τ^2 большое) $\Rightarrow$ модель гибкая, риск переобучения.
- Большое λ (сильный априор, τ^2 маленькое) $\Rightarrow$ коэффициенты сжаты к нулю, риск недообучения.
- Evidence $P(y \mid X, \lambda)$ балансирует эти два эффекта:
  - штрафует за избыточную сложность (бритва Оккама),
  - поощряет модели, которые хорошо объясняют данные.
- В итоге получаем «оптимальное» λ без кросс-валидации.