Обучение с учителем. Байесовский подход

Санкт-Петербургский государственный университет
Кафедра статистического моделирования

20 сентября 2025

Теорема Байеса: ядро подхода

Фундаментальное правило обновления убеждений

В байесовском подходе параметры модели θ рассматриваются как случайные величины. Наши знания о них обновляются при поступлении данных D с помощью теоремы Байеса.

$$\underbrace{p(\theta|D)}_{\text{Апостериор}} = \frac{\overbrace{p(D|\theta)}^{\text{Ф. правдоподобия}} \cdot \overbrace{p(\theta)}^{\text{Априор}}}{\underbrace{p(D)}_{\text{Ф. предельного правдоподобия}}} = \frac{p(D|\theta) \cdot p(\theta)}{\int p(D|\theta) \cdot p(\theta) d\theta}$$

- **Априорное распределение** $p(\theta)$: наши убеждения о параметрах *до* получения данных.
- **Функция правдоподобия** $p(D|\theta)$: вероятность наблюдать данные D при фиксированных параметрах θ.
- **Апостериорное распределение** $p(\theta|D)$: обновленные убеждения о θ *после* учета данных.

Сравнение подходов: частотный и байесовский

Частотный подход

- Вероятность — это долгосрочная частота события.
- Параметры модели (θ) — это фиксированные, неизвестные константы.
- Результат — точечные оценки (например, ОМП) и доверительные интервалы для θ.
- Доверительный интервал (например, 95%) означает, что при многократном повторении эксперимента 95% таких интервалов будут содержать истинное значение параметра.

Байесовский подход

- Вероятность — это степень уверенности в утверждении.
- Параметры модели (θ) — это случайные величины с распределениями.
- Результат — апостериорное распределение параметров θ.
- Доверительный интервал (например, 95%) означает, что существует 95%-ая вероятность того, что истинное значение параметра находится в этом интервале.

Теорема Бернштейна–фон Мизеса

При определенных условиях регулярности, для больших N апостериорное распределение $p(\theta|D)$ сходится к нормальному распределению:

$$p(\theta|D) \approx \mathcal{N}(\theta \mid \hat{\theta}_{\text{MLE}}, [E_N(\hat{\theta}_{\text{MLE}})]^{-1})$$

Следствие:

- Апостериорное распределение становится гауссианой.
- Его центр $\hat{\theta}_{\text{MLE}}$ является состоятельной оценкой истинного значения θ_0.
- Его дисперсия (неопределенность) уменьшается с ростом N.
- **Апостериорное распределение больше не зависит от выбора априорного при больших N.**

Гибридный подход к классификации

Гибридный подход подразумевает, что мы оцениваем параметры правдоподобия (например, среднее и дисперсию для нормального распределения) как точные значения из данных.

Формальный байесовский подход

1. **Моделирование**: Для каждого класса k мы оцениваем плотность вероятности признаков $p(\mathbf{x}|y=k)$ и априорную вероятность класса $p(y=k)$.
2. **Классификация**: Используем теорему Байеса для вычисления апостериорной вероятности:

$$p(y=k|\mathbf{x}) = \frac{p(\mathbf{x}|y=k)p(y=k)}{p(\mathbf{x})}$$

И выбираем класс, который ее максимизирует.

Наивный байесовский классификатор

Ключевое (“наивное”) предположение

Признаки $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ **условно независимы** при заданном классе $y = k$.

Математически это означает:

$$p(\mathbf{x}|y = k) = p(x_1, \ldots, x_d|y = k) = \prod_{j=1}^{d} p(x_j|y = k)$$

Это предположение резко упрощает задачу. Вместо оценки сложной многомерной плотности $p(\mathbf{x}|y = k)$, нам нужно оценить d одномерных плотностей $p(x_j|y = k)$.

Правило классификации

$$\hat{y} = \arg\max_k \left(\log p(y = k) + \sum_{j=1}^{d} \log p(x_j|y = k) \right)$$

Вариации в зависимости от типа данных

Вид классификатора определяется выбором распределения для функции правдоподобия $p(x_j|y=k)$.

Gaussian NB

Данные: Непрерывные
Модель: $p(x_j|y=k)$ — нормальное распределение $\mathcal{N}(\mu_{jk}, \sigma^2_{jk})$
Пример: Классификация ирисов по длине и ширине лепестков

Multinomial NB

Данные: Дискретные (счетчики)
Модель: $p(x_j|y=k)$ — мультиномиальное распределение
Пример: Классификация текстов. x_j — частота j-го слова в документе

Bernoulli NB

Данные: Бинарные (0/1)
Модель: $p(x_j|y=k)$ — распределение Бернулли
Пример: Анализ спама. $x_j = 1$, если j-е слово есть в письме, и 0 иначе

LDA и QDA

Дискриминантный анализ решают задачу классификации, моделируя распределение данных для каждого класса.

Ключевое предположение

Плотность распределения признаков для каждого класса k является многомерным нормальным распределением:

$$p(\mathbf{x}|y=k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Для удобства вычислений переходим к логарифму апостериорной вероятности:

$$\hat{y} = \arg\max_k \left(\log p(\mathbf{x}|y=k) + \log p(y=k)\right)$$

Подставляя логарифм плотности Гаусса, получаем дискриминантную функцию $\delta_k(\mathbf{x})$:

$$\delta_k(\mathbf{x}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) + \log\pi_k$$

где $\pi_k = p(y=k)$ — априорная вероятность класса.

Квадратичный дискриминантный анализ (QDA)

Предположение QDA

Каждый класс k имеет собственную матрицу ковариаций $\boldsymbol{\Sigma}_k$.

Дискриминантная функция $\delta_k(\mathbf{x})$ является квадратичной функцией от $\mathbf{x}$:

$$\delta_k(\mathbf{x}) = -\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}_k^{-1}\mathbf{x} + \mathbf{x}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| + \log\pi_k$$

- Условие $\delta_k(\mathbf{x}) = \delta_j(\mathbf{x})$ задает квадратичную разделяющую поверхность (эллипсоид, параболоид, гиперболоид).
- QDA — очень гибкий метод, но требует оценки большого числа параметров для каждой матрицы $\boldsymbol{\Sigma}_k$, что может привести к переобучению на малых выборках.

Линейный дискриминантный анализ (LDA)

Предположение LDA

Все классы имеют общую матрицу ковариаций $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$.

Это упрощение сильно меняет дискриминантную функцию. Член $\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}$ становится одинаковым для всех классов и не влияет на $\arg\max_k$.

Линейная дискриминантная функция

$$\delta_k(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log \pi_k$$

- Это линейная функция от $\mathbf{x}$.
- Разделяющая поверхность $\delta_k(\mathbf{x}) = \delta_j(\mathbf{x})$ является гиперплоскостью.
- LDA более устойчив к переобучению, чем QDA, так как оценивается меньше параметров.

Визуализация разделяющих поверхностей

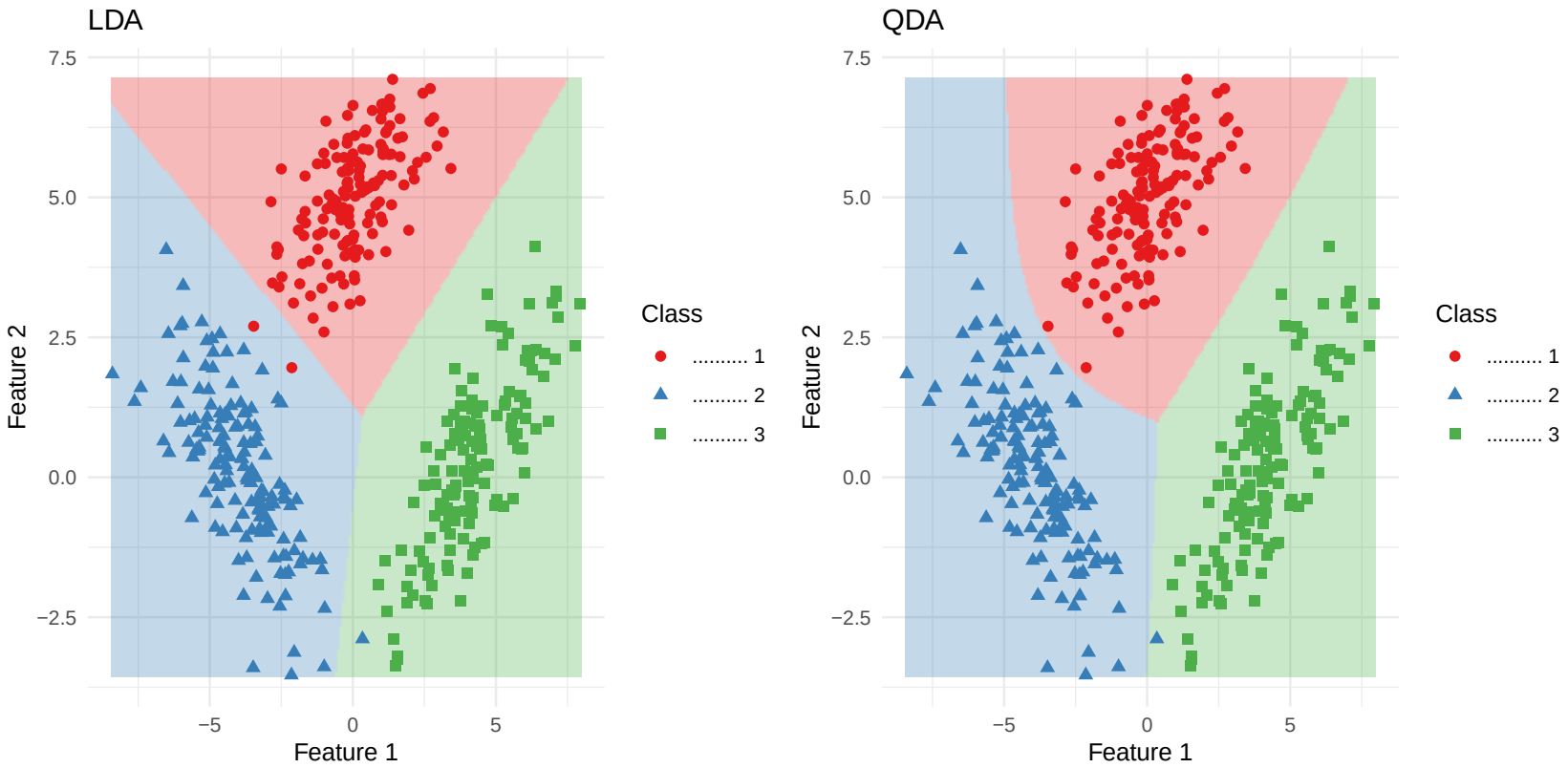

Рис.: Слева: LDA создает линейные границы. Справа: QDA создает гибкие квадратичные границы, которые лучше разделяют классы с разной формой ковариации.

Байесовское иерархическое моделирование

Полностью байесовский (иерархический) подход к классификации заключается в том, что мы не уверены не только в классе объекта, но и в точных параметрах, описывающих этот класс.

Иерархический аспект таких моделей заключается в том, что мы организуем наши параметры на разных уровнях.

Байесовский аспект заключается в том, что мы корректируем наши представления об этих параметрах на основе наблюдаемых данных.

Структура иерархической модели

Параметры модели сами рассматриваются как случайные величины, взятые из распределения более высокого уровня.

- **Уровень 1 (Данные):** $x \sim \text{Распределение}(\theta_k)$
- **Уровень 2 (Параметры классов):** $\theta_k \sim \text{Распределение}(\lambda)$
- **Уровень 3 (Гиперпараметры):** $\lambda \sim \text{Гиперприор}$

Пример: Средние μ_{jk} в гауссовском наивном Байесе

- Вместо того чтобы считать каждое μ_{jk} независимым, мы предполагаем, что все они взяты из общего "материнского" распределения:

$$\mu_{jk} \sim \mathcal{N}(\mu_{j,global}, \sigma^2_{j,global})$$

- Модель оценивает и μ_{jk}, и общие $\mu_{j,global}, \sigma^2_{j,global}$ одновременно.

Оценка для редкого класса становится компромиссом между его собственными данными и средним по всем классам. Это делает модель гораздо более устойчивой к выбросам.

Байесовская линейная регрессия

Частотная линейная регрессия:

- Модель: $y = \mathbf{w}^T \mathbf{x} + \epsilon$, где $\epsilon \sim N(0, \sigma^2)$.
- Цель: найти точечные оценки коэффициентов β, которые минимизируют сумму квадратов ошибок.
- Результат: вектор коэффициентов $\hat{\beta}$ и их доверительные интервалы.

Байесовская линейная регрессия:

- Модель та же, но на параметры $\mathbf{w}$ и σ^2 задаются априорные распределения, например, $\mathbf{w} \sim N(0, \alpha^{-1} I)$.
- Цель: найти апостериорные распределения для $\mathbf{w}$.
- Результат: полные распределения для каждого коэффициента, которые показывают нашу неуверенность в оценках.

Математическая модель

Предполагаем, что шум аддитивный и гауссовский: $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Тогда функция правдоподобия для одного объекта $(\mathbf{x}_i, y_i)$ имеет вид:

$$p(y_i|\mathbf{x}_i, \mathbf{w}) = \mathcal{N}(y_i|\mathbf{w}^T\mathbf{x}_i, \sigma^2)$$

Для всей выборки $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$:
$p(D|\mathbf{w}) = \prod_i \mathcal{N}(y_i|\mathbf{w}^T\mathbf{x}_i, \sigma^2)$.

Априорное распределение

Чтобы избежать переобучения, вводим априорное распределение на веса $\mathbf{w}$. Обычно используется гауссовское распределение с центром в нуле:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\boldsymbol{E})$$

- Параметр α контролирует "разброс" весов. Большое α "стягивает" веса к нулю.
- Это эквивалентно L2-регуляризации в частотном подходе.

Связь с L2-регуляризацией

$$\mathbf{w}_{\text{MAP}} = \arg\max_{\mathbf{w}} p(\mathbf{w}|D) = \arg\max_{\mathbf{w}} \left(\log p(D|\mathbf{w}) + \log p(\mathbf{w})\right)$$

$$\log p(D|\mathbf{w}) = \log \prod_{i=1}^{N} \mathcal{N}(y_i|\mathbf{w}^T\mathbf{x}_i, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2 + \text{const}$$

$$\log p(\mathbf{w}) = \log \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \log\left(C \cdot \exp\left(-\frac{1}{2}\mathbf{w}^T(\alpha^{-1}\mathbf{I})^{-1}\mathbf{w}\right)\right)$$

$$= -\frac{\alpha}{2}||\mathbf{w}||_2^2 + \text{const}$$

$$\mathbf{w}_{\text{MAP}} = \arg\max_{\mathbf{w}} \left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2 - \frac{\alpha}{2}||\mathbf{w}||_2^2\right)$$

$$= \arg\min_{\mathbf{w}} \left(\sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2 + \sigma^2\alpha_\lambda||\mathbf{w}||_2^2\right)$$

Апостериорное распределение для весов

Так как априорное распределение и функция правдоподобия являются гауссовскими, то и апостериорное распределение $p(\mathbf{w}|D)$ тоже будет гауссовским (свойство сопряженности).

$$p(\mathbf{w}|D) \sim \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \boldsymbol{S}_N)$$

где $\mathbf{m}_N$ и $\boldsymbol{S}_N$ вычисляются на основе данных.

Для нового объекта $\mathbf{x}_*$ предсказание y_* — это не одно число, а целое распределение. Оно получается путем усреднения по всем возможным весам $\mathbf{w}$ с учетом их апостериорной вероятности:

$$p(y_*|\mathbf{x}_*, D) = \int p(y_*|\mathbf{x}_*, \mathbf{w}) p(\mathbf{w}|D) d\mathbf{w}$$

Это распределение также является гауссовским. Его среднее — это наше предсказание, а дисперсия — мера нашей неуверенности в этом предсказании.

Вычисление апостериорного распределения

Для вычисления апостериорного распределения $p(\theta|D) \propto p(D|\theta)p(\theta)$ необходимо вычислить знаменатель (функцию предельного правдоподобия):

$$P(D) = \int P(D|\theta)P(\theta)d\theta$$

- Этот интеграл часто является невычислимым аналитически, особенно в моделях с большим числом параметров.
- Существует два основных подхода к решению этой проблемы:
  1. Аналитический вывод с использованием сопряженных априорных распределений.
  2. Численные методы, такие как методы Монте-Карло по схеме марковских цепей (MCMC).

Аналитический подход: Сопряженные распределения

Определение

Априорное распределение $P(\theta)$ называется сопряженным для функции правдоподобия $P(D|\theta)$, если получаемое апостериорное распределение $P(\theta|D)$ принадлежит тому же семейству распределений, что и априорное.

Вместо сложного интегрирования, весь процесс байесовского обновления сводится к простому вычислению новых параметров по готовым формулам. Это делает вывод быстрым и аналитически разрешимым. Выбор сопряженных пар (Нормальное–Нормальное, Бета–Бернулли, Гамма–Пуассона и т.д.) — это основной способ сделать байесовский анализ вычислительно возможным.

Пример:

- **Правдоподобие (Биномиальное):** Моделируем k успехов в n испытаниях с вероятностью успеха θ.

$$P(D|\theta) = \text{Bin}(k|n,\theta) \propto \theta^k(1-\theta)^{n-k}$$

- **Априорное (Бета):** Предполагаем, что θ следует Бета-распределению.

$$P(\theta) = \text{Beta}(\theta|\alpha,\beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- **Апостериорное (Бета):** Апостериорное распределение также является Бета-распределением.

$$P(\theta|D) \propto \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1} = \text{Beta}(\theta|k+\alpha, n-k+\beta)$$

Численный подход: Методы MCMC

- **Идея:** Если мы не можем аналитически описать апостериорное распределение, мы можем сгенерировать из него выборку.
- MCMC (Markov Chain Monte Carlo) — это алгоритм, строящий марковскую цепь, стационарное распределение которой совпадает с искомым апостериорным распределением $p(\theta|D)$.
- Использование сопряженных априорных распределений удобно, но часто является сильным упрощением.
- Мы хотим использовать более гибкие или более реалистичные априорные распределения, которые не являются сопряженными.
- Модель сложна, и найти сопряженное распределение для всех параметров невозможно.

Пример:

1. Начинаем со случайного значения параметра θ_0.
2. На каждой итерации t:
   - Предлагаем нового кандидата θ' из некоторого предложенного распределения $q(\theta'|\theta_{t-1})$.
   - Вычисляем коэффициент принятия $\alpha = \frac{p(\theta'|D)}{p(\theta_{t-1}|D)}$. Так как $p(\theta|D) \propto p(D|\theta)p(\theta)$, нам не нужно знать нормализующую константу.
   - Принимаем кандидата $(\theta_t = \theta')$ с вероятностью $\min(1, \alpha)$, иначе оставляем старое значение $(\theta_t = \theta_{t-1})$.
3. После достаточного числа итераций полученные значения $\{\theta_t\}$ будут выборкой из апостериорного распределения.