

Санкт-Петербургский государственный университет
Прикладная математика и информатика

Отчет по производственной практике (педагогической)

БАЙЕСОВСКОЕ ОБУЧЕНИЕ (СЕМЕСТР 3)

Выполнил:

Дудник Павел Дмитриевич

группа 23.М03-мм

Научный руководитель:

д. ф.-м. н., доцент

Голяндина Нина Эдуардовна

Кафедра Статистического Моделирования

Оглавление

Введение	3
1. Теоретическая основа	3
1.1. Теорема Байеса	3
1.2. Априорные и апостериорные распределения	3
1.3. Композитность	4
2. Байесовское обучение с информативными априорными распределениями	5
3. Байесовское обучение без использования информации о параметре	7
3.1. Работа с апостериорными распределениями	7
3.2. Предсказательное распределение	8
3.3. Апостериорный предсказательный p-value	8
4. Пример обновления весов	9
5. Заключение	13
Список литературы	14

Введение

Байесовское обучение [1, 2] — это вероятностный подход к статистическому выводу, основанный на теореме Байеса. В отличие от классического метода максимального правдоподобия (ММП), байесовский подход позволяет учитывать априорные знания о параметрах модели и предоставляет распределение их неопределённости. Это позволяет сделать получение оценок параметров итеративным, с получением новых данных.

Педагогическая задача заключается в том, чтобы ознакомить студентов с понятием байесовского обучения в достаточном объеме, привести примеры задач.

1. Теоретическая основа

1.1. Теорема Байеса

Базовый инструмент байесовского подхода — теорема Байеса:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)},$$

где:

- $P(\theta)$ — априорное распределение параметров,
- $P(D|\theta)$ — правдоподобие данных,
- $P(\theta|D)$ — апостериорное распределение.

1.2. Априорные и апостериорные распределения

Рассмотрим пример. Рассмотрим задачу с подбрасыванием монетки, где параметр θ — это вероятность выпадения орла. Пусть априорное распределение θ равномерное на интервале $[0; 1]$. Мы хотим обновить апостериорное распределение θ после наблюдения данных. Правдоподобие данных описывается биномиальным распределением:

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

т.е.

$$P(D|\theta) \propto \theta^k (1 - \theta)^{n-k}.$$



Рис. 1. Переход к апостериорному распределению.

Таким образом,

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

$$P(\theta|D) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

Результатом решения задачи в байесовском подходе является апостериорное распределение. Отметим, что и для ММП можно получать оценки из распределений, достаточно посчитать стандартное отклонение параметра. Таким образом, байесовское обучение заключается в гибкой инкрементальной работе с данными для получения оценок.

1.3. Композитность

Композитность — возможность объединить знания:

$$p(y|x, z) = \frac{p(z|y)p(y|x)}{\int p(z|y)p(y|z)dz}$$

Модель задается как параметрическое семейство распределений $\{\mathbb{P}_\theta | \theta \in \Theta \in \mathbb{R}^k\}$ с заданным на Θ априорным распределением $p(\theta)$.

Каждый раз при получении новых данных мы обновляем наше апостериорное распределение.

Пусть мы m раз обновляли апостериорное распределение, получая данные x_i :

$$p(\theta) \rightarrow p(\theta|x_1) \rightarrow \rightarrow p(\theta|x_1, x_2) \rightarrow \cdots \rightarrow p(\theta|x_1, \dots, x_m)$$

Понятно, что m -ый шаг равносильно одному шагу с данными $x_1, \dots, x_m$.

Что будет происходить с апостериорным распределением, если будет поступать неограниченное количество данных?

- При выполнении *условий регулярности*, апостериорное распределение $p(\theta|x_m)$ при $m \rightarrow \infty$ приближается к нормальному распределению $\mathcal{N}(\theta_0, (mI(\theta_0))^{-1})$.
- θ_0 — параметр истинного распределения, если распределение лежит в нашей модели.
- Если распределение не лежит в нашей модели, то θ_0 минимизирует KL-расстояние от истинного распределения до нашей модели.
- $I(\theta_0)$ — информация Фишера.

Т.е. $p(\theta|x_m)$ стремится к вырожденному распределению в точке θ_0 .

Что если θ — многомерный параметр?

Пусть есть апостериорное распределение $p(\theta|x)$. По нему мы можем посчитать **маргинальное распределение** конкретной компоненты θ или группы компонент:

$$p(\theta_i|x) = \int p(\theta|x) d\theta_{-i}$$

Такие интегралы часто приходится считать численно.

2. Байесовское обучение с информативными априорными распределениями

Байесовское обучение может встречаться не в контексте технической реализации вычисления параметра, а в контексте наличия важной информации о распределении параметра. Например, имеются результаты экспериментов, при которых достоверный интервал для параметра небольшой.

Априорные распределения обычно разделяют на следующие виды [3]:

Виды априорных распределений

Пусть $d = [1,1,1,1,1,1,1,1,1,1]$ — десять бросков монетки.

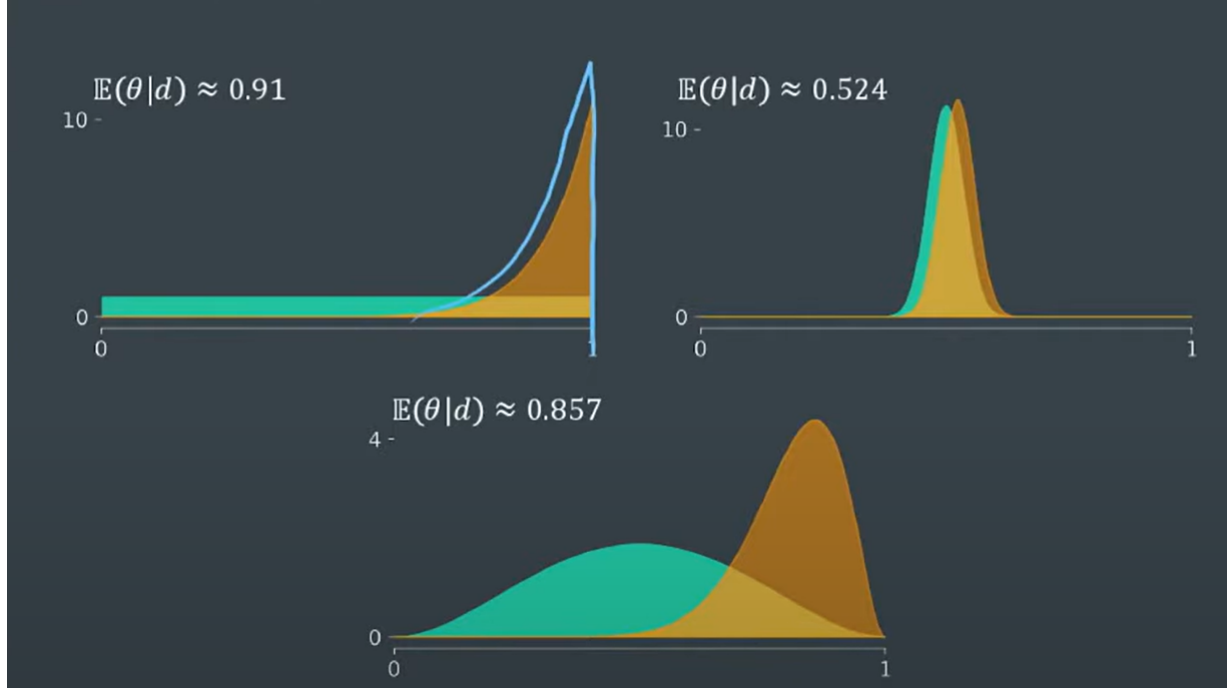


Рис. 2. Априорные и апостериорные распределения.

- **Неинформативные:** не несут дополнительной информации о значениях параметров (равномерные на θ). Подходят для случая, когда мы не можем предпочесть конкретное значение любому другому.
- **Информативные:** несут существенную дополнительную информацию о значениях параметра. Например: накопленные знания в некоторой предметной области. Если данных мало, то априорное распределение может ухудшить понимание.
- **Слабо информативные:** несут частичную информацию. Избавляемся от экстремальных значений. Такой выбор можно назвать регуляризацией для параметра.

3. Байесовское обучение без использования информации о параметре

Когда нет дополнительной информации о параметре, байесовское обучение можно рассматривать как технический прием - итеративное вычисление параметра с поступлением новых данных.

3.1. Работа с апостериорными распределениями

Иногда описать апостериорное распределение полностью невозможно. Что можно использовать вместо него?

1. Точечные характеристики: среднее $\int \phi(\theta)p(\theta|x)d\theta$, максимум $\operatorname{argmax}_{\phi(\theta)}p(\theta|x)$.
2. **Достоверный интервал** — множество, в котором $\phi(\theta)$ лежит с определенной вероятностью.

Если данных много, то искать апостериорное распределение может быть затруднительно: $\int p(x|\theta)p(\theta)d\theta$ не считается аналитически для любых $p(x|\theta)$ и $p(\theta)$.

Что делать в таком случае?

Сопряженное апостериорное распределение — распределение из параметрического семейства, для которого априорное распределение также в этом семействе. Таким образом, часто априорное распределение выбирают исходя из наличия сопряженного апостериорного.

Примеры сопряженных распределений:

- Бернулли - Гамма
- Биномиальное - Гамма
- Нормальное с изв. σ^2 - Нормальное (параметр μ)
- Нормальное с изв. μ - Гамма (параметр σ^2)

Пример:

- $X \sim B(\theta) \Rightarrow p(x_n|\theta) = \theta^{\sum x_i} (1 - \theta)^{(n - \sum x_i)}$.

- Априорное распределение: $\theta \sim \text{Beta}(\alpha, \beta)$, $p(\theta) = \theta^{\alpha-1}(1-\theta)^{\beta-1}/B(\alpha, \beta)$.
- $p(\theta|x_n) \propto \theta^{\sum x_i + \alpha - 1}(1-\theta)^{n - \sum x_i + \beta - 1}$
- Получаем, что $\theta|x_n \sim \text{Beta}(\alpha + \sum x_i, \beta + n - \sum x_i)$.
- Таким образом, при получении новых данных, для получения апостериорного распределения, нужно будет вычислить параметры Beta распределения.

3.2. Предсказательное распределение

Пусть мы умеем считать апостериорное распределение $p(\theta|X)$. Тогда мы можем посчитать $p(X_{new}|X)$ — **предсказательное распределение**.

$$\begin{aligned} p(X_{new}|X) &= \int_{\theta} p(X_{new}, \theta|X) d\theta = \int_{\theta} p(X_{new}|\theta, X) p(\theta|X) d\theta \\ &= \int_{\theta} p(X_{new}|\theta) p(\theta|X) d\theta \end{aligned}$$

Однако на практике проще генерировать данные следующим образом:

1. Генерируем θ' из апостериорного распределения.
2. Генерируем новое наблюдение X_{new} из $p(*|\theta')$.

3.3. Апостериорный предсказательный p-value

Оказывают ли недостатки модели существенное влияние на выводы?

- Если модель хорошая, то данные из предсказательного распределения должны быть похожи на реальные данные.
- Похожесть можно определять с помощью тестовых статистик T .

$P(T(X_{new}) > T(X)|X)$ — **апостериорный предсказательный p-value**.

- Если апостериорный предсказательный p-value близок к 0 или 1, то качество модели низкое.

Как считать апостериорный предсказательный p-value?

$$p = \int \int [T(X_{new}) \geq T(X)] p(X_{new}|\theta) p(\theta|X) d(X_{new}) d\theta$$

Как считают на самом деле:

1. Генерируют K параметров θ' из апостериорного распределения.
2. Для каждого из них генерируют $X_{new} \sim p(*|\theta')$.
3. Для каждого X_{new} считают $T(X_{new})$.
4. Значение p примерно равно доле $T(X_{new})$ больших $T(X)$.

4. Пример обновления весов

Рассмотрим пример с линейной регрессией на языке программирования Python.

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn import metrics

# Генерация синтетических данных
np.random.seed(42)
n_samples = 200
X = np.linspace(-10, 10, n_samples)
np.random.shuffle(X)
w_actual = 3.5 # Истинный вес
b_actual = -4. # Истинное смещение
y = X * w_actual + b_actual
y_noised = y + np.random.normal(0, 1, n_samples) # Добавляем шум

# Визуализация данных
plt.scatter(X, y_noised, label="Данные")
plt.plot(X, y, color='red', label="Истинная модель")
plt.xlabel("X")
plt.ylabel("y")
plt.legend()
plt.title("Синтетические данные")
plt.savefig("data.png")
plt.show()
```

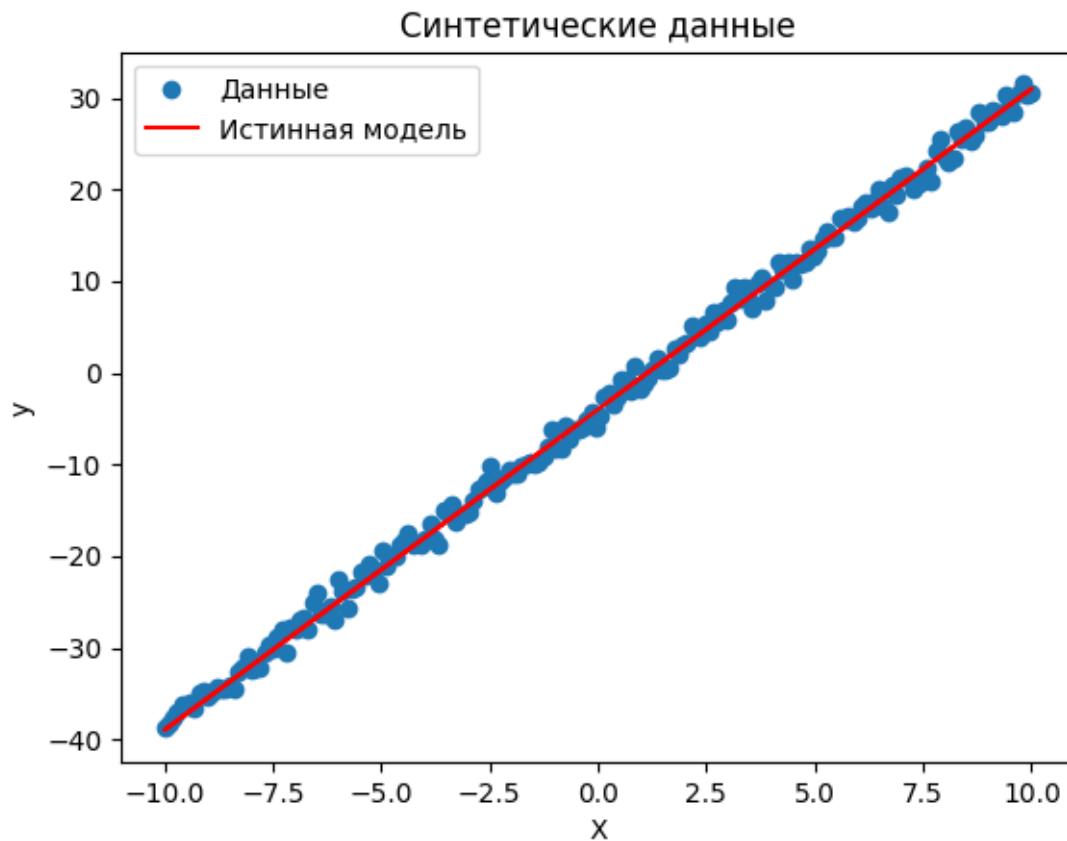


Рис. 3. Сгенерированные данные.

```
X_b = np.concatenate(
    [X.reshape(-1, 1),
     np.ones_like(X).reshape(-1, 1)],
    axis = 1) # столбец с 1 для смещения

# Разделим на тренировочные и тестовые данные
X_train, X_test, Y_train, Y_test = train_test_split(
    X_b, y_noised, test_size=100, random_state=42)

# Разделим данные на батчи
batch_size = 10
num_batches = int(np.ceil(X_train.shape[0] / batch_size))
X_train = np.array_split(X_train, num_batches)
Y_train = np.array_split(Y_train, num_batches)
```

```

# Байесовское обновление весов
# Априорное распределение:  $w \sim N(0, \alpha^{-1} I)$ ,  $\alpha = 1$ .
alpha = 1.0

# Априорная ковариация и среднее для весов
prior_cov = np.eye(2) / alpha # Ковариация априорного распределения
prior_mean = np.zeros(2) # Среднее априорного распределения

def update_posterior(X, y, prior_mean, prior_cov):
    # Апостериорная ковариация
    posterior_cov = np.linalg.inv(np.linalg.inv(prior_cov) + (X.T @ X))

    # Апостериорное среднее
    posterior_mean = posterior_cov @ (np.linalg.inv(prior_cov) @ \
                                      prior_mean + (X.T @ y))

    return posterior_mean, posterior_cov

for X_batch, Y_batch in zip(X_train, Y_train):
    posterior_mean, posterior_cov = update_posterior(
        X_batch, Y_batch, prior_mean, prior_cov)

    print('posterior mean: ')
    print(posterior_mean)
    print('posterior cov: ')
    print(posterior_cov)

    prior_mean = posterior_mean
    prior_cov = posterior_cov

    pred = X_test @ prior_mean.reshape(-1, 1)
    mse = metrics.mean_squared_error(Y_test, pred)
    print(f'Test MSE: {mse}')

n_samples_posterior = 100

```

```

w_samples = np.random.multivariate_normal(
    posterior_mean, posterior_cov, n_samples_posterior)

# Визуализация
plt.figure(figsize=(10, 6))
data = plt.scatter(X_test[:,0], Y_test, label="Данные", alpha=0.5)
for w, b in w_samples:
    plt.plot(X_test[:,0], w * X_test[:,0] + b, color='green', alpha=0.1)
model, = plt.plot(X_test[:,0], w_actual * X_test[:,0] + b_actual, color='red',
    label="Истинная модель")
plt.xlabel("X")
plt.ylabel("y")
empty = plt.Line2D([], [], color='green', label='Полученные модели')
plt.legend(handles=[data, model, empty])
plt.title("Апостериорное распределение весов")
plt.savefig("result.png")
plt.show()

```

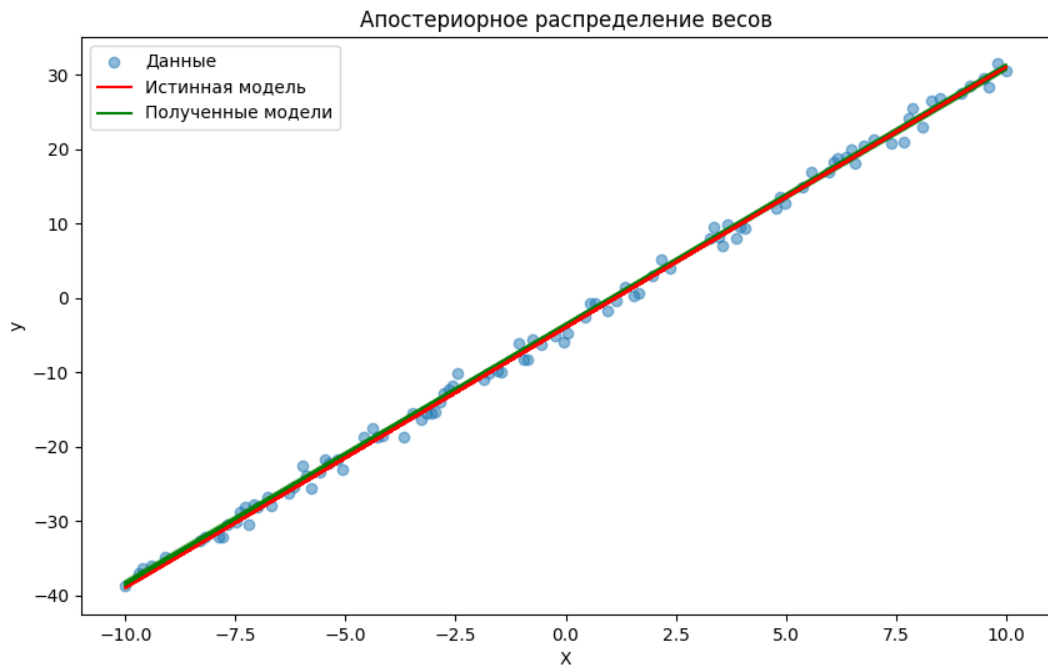


Рис. 4. Предсказания моделей с параметрами из апостериорного распределения.

5. Заключение

Байесовское обучение представляет собой гибкий подход к статистическому выводу, который позволяет учитывать априорные знания о параметрах модели и оценивать их неопределённость через апостериорное распределение. Главные отличия от ММП — это возможность инкрементального обновления параметров. Однако вычислительная сложность байесовского подхода может быть ограничением для задач с большими данными. В таких случаях ММП остаётся более практичным выбором.

Выполненные педагогические задачи: описано понятие байесовского обучения, приведены примеры задач.

Список литературы

1. Murphy K. P. Machine learning: A Probabilistic Perspective. — 2012.
2. Gelman A. Bayesian Data Analysis. — 2013.
3. Nikolaev M. Bayesian approach. — 2022. — Access mode: <https://www.youtube.com/live/Hv3Nw94Q32s?si=cZeWzo4oSYUWmcUu> (online; accessed: 2025-03-07).