EPFL, Spring 2023
Guillaume Obozinski

EE-209 Eléments de Statistiques pour les Data Sciences

Feuille d'exercices 9

Exercise 9.1 Airline accidents.

We are interested in the number of airline accidents in the US per year. The dataset is available on the website of *Airlines for America*[1]. Our goal is to use the Bayesian paradigm to predict the number of airline accidents in the US per year. Since the data available is the yearly number of accidents since the year 2000, we will denote by Y_t the random variable giving the number of accidents during year $2000+t$, with t an integer ≥ 0. We assume that the $(Y_t)_t$ are i.i.d. and follow a Poisson distribution with parameter $\theta > 0$. Let $\mathcal{T} = \{0, \ldots, 19\}$ the indices of y corresponding to the years from 2000 to 2019 We will denote $n = |\mathcal{T}| = 20$ the number of considered years

(a) We consider a Gamma prior with parameter shape parameter α and with rate parameter β for the Poisson parameter θ. We denote this prior by $\mathcal{G}(\alpha, \beta)$ and we recall that its pdf $p(\theta)$ is such that

$$p(\theta) \propto \theta^{\alpha-1} \, e^{-\beta\theta} \, 1_{\{\theta>0\}}.$$

Compute the posterior distribution of θ given a single observation y_0.

(b) Prove that the posterior distribution of θ given $(y_t)_{t\in\mathcal{T}}$ is still a Gamma distribution and give its shape and rate parameter.

(c) Using the previous question, deduce the posterior mean of θ. Show that the posterior mean of θ can be written as

$$(1-\gamma_n)\,\widehat{\theta}_{\text{MLE}} + \gamma_n\,\theta_0,$$

where $\gamma_n \in (0,1)$, $\widehat{\theta}_{\text{MLE}}$ is the MLE of the parameter θ and θ_0 is the prior mean.

(d) Given the decomposition of the posterior mean obtained in the previous question, it is possible to interpret the shape parameter α and the rate parameter β of the prior distribution as representing pseudo-observations and/or pseudo-counts?

Exercise 9.2

We consider the usual Gaussian model $\mathcal{N}(\mu, \sigma^2)$, and given a sample of observed data $\mathcal{D}_n = \{x_1, \ldots, x_n\}$, we consider the Bayesian posterior mean estimator of the parameter μ assuming that σ^2 is known. We derived the form of this posterior mean in class which is:

$$\mu_{\text{post}} = \frac{n\tau^2}{n\tau^2+\sigma^2}\overline{X} + \frac{\sigma^2}{n\tau^2+\sigma^2}\mu_0$$

[1] `https://www.airlines.org/dataset/safety-record-of-u-s-air-carriers/`

Since this is a *point estimator*, it is possible to consider and analyze this estimator from a frequentist standpoint. In particular, from that point of view natural questions are what are the usual bias and variance of the estimator, under the distribution of the data. (Note that since this is the posterior mean, the Bayesian uncertainty has been integrated out, and what we would like to analyse in this exercise is not about the posterior distribution itself, which has been entirely summarized as a single value: the posterior mean).

We assume that the data follows a distribution whose expectation is μ (which is unknown) and whose variance is indeed σ^2 (which is assumed known).

(a) What is the bias $\text{Bias}(\mu_{\text{post}}, \mu)$ of this estimator?

(b) What is the variance of this estimator?

(c) Deduce from the previous question the mean square error of μ_{post}. To simplify notations you can introduce $\gamma_n := \frac{\sigma^2}{n\tau^2+\sigma^2}$.

(d) Deduce from the previous question that μ_{post} is a consistent estimator. Would it have been possible to see this directly from the form of μ_{post}?

(e) We now would like to know whether μ_{post} can outperform $\overline{X}$ in MSE. Compute the difference $\text{MSE}(\overline{X}, \mu) - \text{MSE}(\mu_{\text{post}}, \mu)$ and determine under which condition this quantity is positive.

(f) What are the conditions for $\text{MSE}(\mu_{\text{post}}, \mu)$ to be much smaller than $\text{MSE}(\overline{X}, \mu)$.

Exercise 9.3 (Optional) Plotting priors and posteriors for the Airline accident problem in Python. This exercice is entirely optional, for those of you who would want to visualize the prior and posterior distributions from the exercise on Airline accidents.

(a) Using the code provided below (filling the dots to define the variables *alpha_post* and *beta_post* corresponding the the parameters of the Gamma distribution of the posterior), plot on the same figure the prior and the posterior distributions of θ. We consider $\beta = 1$ for your Gamma prior and we take α so that the prior mean is equal to the empirical mean of the data.

```python
import os
import pandas as pd
import numpy as np
from scipy.stats import gamma, poisson, nbinom
import matplotlib.pyplot as plt

# Data loading
nb_accidents = [49, 41, 34, 51, 23, 34, 26, 26, 19, 26, 28, 29,
    23, 18, 29, 27, 26, 30, 28, 36, 10]
years = [2000+i for i in range(len(nb_accidents))]

df = pd.DataFrame(data={'Year':years, 'Accidents':nb_accidents
  })
df = df.set_index('Year')
```

```
# We show that the rate of road accidents in the US between 1994 and 2020
df.plot.bar()
plt.ylabel('Number␣of␣accidents', fontsize=14)
plt.xlabel('Year', fontsize=14)
plt.show()

# We keep only the data before 2019
data = df.filter(items = range(2000,2020), axis=0)

# We plot the prior and the posterior distributions
mean = data['Accidents'].mean()

beta = 1
alpha = mean * beta

print('The␣parameters␣for␣the␣prior␣distribution␣are␣alpha={0}␣and␣beta={1}'.format(round(alpha,3), round(beta,3)))

x = np.linspace(0, 40, 100)
prior = gamma.pdf(x, a=alpha, scale=1/beta)


#### TO COMPLETE: using the command 'data['Accidents'].sum()' gives the sum of the number of accidents between 2000 and 2019
alpha_post = .....
beta_post  = .....


posterior = gamma.pdf(x, a=alpha_post, scale=1/beta_post)

plt.plot(x, prior, label='prior')
plt.plot(x, posterior, label='posterior')
plt.legend(fontsize=14)
plt.ylabel('Accidents', fontsize=14)
plt.savefig('prior_posterior.png', dpi=250)
plt.show()
```