MACHINE LEARNING FUNDAMENTALS - LS2026
SEMINAR: BAYESIAN LEARNING

CZECH TECHNICAL UNIVERSITY IN PRAGUE
V. FRANC

Assignment 1. Consider a small training set

$$T_m = (x_i \in \{0,1\} \mid i = 1,\dots,m),$$

where the observations are assumed to be independent and identically distributed according to a Bernoulli distribution

$$p(x_i \mid \theta) = \theta^{x_i}(1-\theta)^{1-x_i}.$$

The goal is to infer the unknown θ. We adopt a Bayesian approach and choose a Beta prior distribution for θ:

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

with hyperparameters $\alpha > 0$ and $\beta > 0$.

a) Using Bayes' rule, show that the posterior distribution

$$p(\theta \mid T_m) \propto p(T_m \mid \theta)p(\theta)$$

is also a Beta distribution. Identify its parameters.

b) Derive the posterior predictive probability of observing $x = 1$, namely

$$p(x = 1 \mid T_m).$$

Use the fact that mean of the Beta distribution reads

$$\mathbb{E}_{\theta \sim Beta(\theta|\alpha,\beta)}[\theta] = \frac{\alpha}{\alpha+\beta}\,.$$

Solution 1. Let

$$S = \sum_{i=1}^{m} x_i$$

denote the number of observations equal to 1. Since $x_i \in \{0,1\}$, the number of observations equal to 0 is $m - S$.

a) The likelihood of the training set is

$$p(T_m \mid \theta) = \prod_{i=1}^{m} p(x_i \mid \theta).$$

Using the Bernoulli model,

$$p(T_m \mid \theta) = \prod_{i=1}^{m} \theta^{x_i}(1-\theta)^{1-x_i}.$$

Therefore,

$$p(T_m \mid \theta) = \theta^{\sum_{i=1}^{m} x_i}(1-\theta)^{\sum_{i=1}^{m}(1-x_i)}.$$

Since

$$\sum_{i=1}^{m} x_i = S \qquad \text{and} \qquad \sum_{i=1}^{m}(1-x_i) = m - S,$$

we get

$$p(T_m \mid \theta) = \theta^S (1-\theta)^{m-S}.$$

The prior distribution is

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

By Bayes' rule,

$$p(\theta \mid T_m) \propto p(T_m \mid \theta)p(\theta).$$

Substituting the likelihood and the prior gives

$$p(\theta \mid T_m) \propto \theta^S (1-\theta)^{m-S}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

Combining powers of θ and $1-\theta$, we obtain

$$p(\theta \mid T_m) \propto \theta^{\alpha+S-1}(1-\theta)^{\beta+m-S-1}.$$

This has the form of a Beta distribution. Hence,

$$p(\theta \mid T_m) = \mathrm{Beta}(\theta \mid \alpha + S, \beta + m - S).$$

Therefore, the posterior parameters are

$$\alpha' = \alpha + S$$

and

$$\beta' = \beta + m - S.$$

b) The posterior predictive probability of observing $x = 1$ is

$$p(x = 1 \mid T_m) = \int_0^1 p(x = 1 \mid \theta)p(\theta \mid T_m)\, d\theta.$$

Since

$$p(x = 1 \mid \theta) = \theta,$$

we have

$$p(x = 1 \mid T_m) = \int_0^1 \theta\, p(\theta \mid T_m)\, d\theta.$$

This is the posterior expectation of θ:

$$p(x = 1 \mid T_m) = \mathbb{E}[\theta \mid T_m].$$

Because

$$\theta \mid T_m \sim \mathrm{Beta}(\alpha + S, \beta + m - S),$$

and the mean of a Beta distribution $\mathrm{Beta}(a, b)$ is

$$\frac{a}{a+b},$$

we get

$$p(x = 1 \mid T_m) = \frac{\alpha + S}{(\alpha + S) + (\beta + m - S)}.$$

Thus,

$$p(x = 1 \mid T_m) = \frac{\alpha + \sum_{i=1}^m x_i}{\alpha + \beta + m}.$$

Assignment 2. Given a small training set $T_m = (x_i \in \mathbb{R} \mid i = 1, \dots, m)$ we want to estimate the mean of a normal distribution $\mathcal{N}(\mu, 1)$. We know that the unknown μ is close to μ_0. Therefore, we want to apply Bayesian inference and set the prior distribution for μ to be a normal distribution centred at μ_0, i.e. $p(\mu) = \mathcal{N}(\mu_0, 1)$.

Show that the posterior distribution $p(\mu \,|\, T_m) \propto p(T_m \,|\, \mu)\, p(\mu)$ is also a Gaussian. Find its center (expectation).

Solution 2. 1. We write down likelihood and prior. The likelihood, up to a constant not depending on μ, reads:

$$p(T_m \mid \mu) \propto \prod_{i=1}^m \exp\left(-\frac{1}{2}(x_i - \mu)^2\right) = \exp\left(-\frac{1}{2}\sum_{i=1}^m (x_i - \mu)^2\right).$$

The prior:

$$p(\mu) \propto \exp\left(-\frac{1}{2}(\mu - \mu_0)^2\right).$$

Posterior (up to proportionality):

$$p(\mu \mid T_m) \propto p(T_m \mid \mu)\, p(\mu) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^m (x_i - \mu)^2 - \frac{1}{2}(\mu - \mu_0)^2\right).$$

We can further write:

$$\begin{aligned}
-\frac{1}{2}\sum_{i=1}^m (x_i - \mu)^2 - \frac{1}{2}(\mu - \mu_0)^2 &= -\frac{1}{2}\left[\sum_i (x_i^2 - 2x_i\mu + \mu^2) + \mu^2 - 2\mu\mu_0 + \mu_0^2\right] \\
&= -\frac{1}{2}\left[\sum_i x_i^2 - 2\mu\sum_i x_i + m\mu^2 + \mu^2 - 2\mu\mu_0 + \mu_0^2\right] \\
&= -\frac{1}{2}\left[\sum_i x_i^2 + \mu_0^2 - 2\mu\big(\sum_i x_i + \mu_0\big) + (m+1)\mu^2\right] \\
&= -\frac{1}{2}(m+1)\left[\frac{\sum_i x_i^2 + \mu_0^2}{m+1} - 2\mu\frac{\sum_i x_i + \mu_0}{m+1} + \mu^2\right] = (*)
\end{aligned}$$

Let us define

$$\hat{\mu} = \frac{\sum_i x_i + \mu_0}{m+1}$$

and write

$$(\mu - \hat{\mu})^2 = \mu^2 - 2\mu\hat{\mu} + \hat{\mu}^2.$$

Hence

$$(*) = \frac{1}{2}(m+1)\left[(\mu - \hat{\mu})^2 \underbrace{- \hat{\mu}^2 + \frac{\sum_i x_i^2 + \mu_0^2}{m+1}}_{\text{const not depending on } \mu}\right]$$

and therefore

$$p(\mu \mid T_m) \propto \exp\left(\frac{1}{2}(m+1)(\mu - \hat{\mu})^2\right) .$$

We have shown that $p(\mu \mid T_m)$ is Gaussian with:

- mean $\hat{\mu} = \frac{\sum_i x_i + \mu_0}{m+1} = \frac{m\bar{\mu} + \mu_0}{m+1}$ where $\bar{\mu} = \frac{1}{m}\sum_i x_i$, and
- variance $\frac{1}{m+1}$.