Weekly Overview Slides of Statistical Machine Learning CSE 575, Fall 2023

Moses A. Boudourides[1]

SPA and SCAI
Arizona State University

[1] Moses.Boudourides@asu.edu

Week 3

Inferring Probability Models from Data

January 26, 2023

Agenda

- About the Project Proposals
- Reference material on Useful Probability Distributions, Chapter 5 of Forsyth's book
- Presentation and discussion of how to infer probability models from data, based on Chapter 9 of Forsyth's book

Inferring Probability Models from Data

- Given a dataset,
  - First, **fit a probability model** to the dataset.
  - Subsequently, **predict or estimate** properties on future new data.
- Potentially, **any model can be fitted** on a dataset (though inefficiently), **even different from the model that produced the dataset**!
- Thus, given a dataset and a probability model that one believes applies to one's dataset, one needs to **estimate the values for the model parameters**.
- For the purpose of estimating model parameters, in general, there are two procedures:
  - **Maximum likelihood**, which finds the parameter values that make the observed data most likely.
  - **Bayesian inference**, which produces a posterior probability distribution on the parameter values, updating the existing prior distribution, and extracts information from that.

Examples of model parameters estimation

1. Repeated flips of a coin – Binomial distribution.
2. Repeated flips of a coin – Geometrical distribution.
3. Detecting spam email – Poisson distribution.
4. Inferring the mean and the standard deviation of normal (Gaussian) data.

The Inverse Problem

Notation:

- $\mathcal{D}$ denotes a *dataset*.
- θ denotes a *parameter* of a (believed) *probability model* of the data $\mathcal{D}$.
- If we knew θ, the probability of observing the data $\mathcal{D}$ would be computed (and denoted) as $P(\mathcal{D}|\theta)$.

The Question:

- Assume that we know the data $\mathcal{D}$, but we do not know the parameter(s) θ.
- How could we estimate the value of θ from the assumed model probability $P(\mathcal{D}|\theta)$, the latter being a function of θ?

Likelihood

Definition (Likelihood)

The function $P(\mathcal{D}|\theta)$ of the unknown variable θ is called **likelihood** of the data $\mathcal{D}$ and it is often denoted as $\mathcal{L}(\theta)$ (or as $\mathcal{L}(\theta; \mathcal{D})$ in order to remember the data involved).

Definition (Maximum Likelihood Principle)

The **Maximum Likelihood Estimate** (**MLE**) chooses θ in such a way that $\mathcal{L}(\theta) = P(\mathcal{D}|\theta)$ is maximized as a function of θ.

Maximum Likelihood Estimation (MLE)

Procedure (**Maximum Likelihood Estimation (MLE)**)

Given a dataset $\mathcal{D}$ and a probability model with unknown parameter(s) θ, compute an estimate $\hat{\theta}$ of the value of the parameters by finding the the value of $\theta = \hat{\theta}$ which maximizes the likelihood of the data under the model:

$$\mathcal{L}(\theta) = P(\mathcal{D}|\theta).$$

In other words, solve the optimization problem:

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta).$$

Independent and Identically Distributed Data (IID)

A data are said to be **independent and identically distributed data** (**IID**) if each data item is an independently obtained sample from the same probability distribution. In terms of the likelihood function, this means that the likelihood is a product of terms, one for each data item, i.e., it is written as:

$$\mathcal{L}(\theta) = P(\mathcal{D}|\theta) = \prod_{d_i \in \mathcal{D}} P(d_i|\theta).$$

Examples

Coin flips – Binomial distribution

If, in N independent coin flips, k heads (H) are observed, then the MLE of the probability $p(H)$ is:

$$\hat{\theta} = \frac{k}{N}.$$

Multiple rolls – Multinomial distribution

If a die is thrown N times and n_1 ones, ..., n_6 sixes are seen, then, denoting by $p_1, \ldots, p_6$ the probabilities that the die comes up as one, ..., six, the MLE of $p_1, \ldots, p_6$ is:

$$\hat{\theta} = \frac{1}{(n_1 + \ldots + n_6)}(n_1, \ldots, n_6).$$

Log–Likelihood

Definition (**Log–Likelihood of a Dataset Under a Model**)

The **log–likelihood** of a dataset under a model is the following logarithmic function of the probability distribution of the unknown parameters:

$$\begin{aligned}\log \mathcal{L}(\theta) &= \log P(\mathcal{D}|\theta) \\ &= \sum_{d_i \in \mathcal{D}} \log P(d_i|\theta).\end{aligned}$$

Poisson Distributions

N intervals are observed, each one having the same fixed length (in time or space), and in each one some events may occur following a Poisson distribution with the same intensity for each observation. Let n_i be the number of events observed in the i'th interval. Then the MLE of the intensity of the observed Poisson distribution is $\hat{\theta} = \frac{\sum_i n_i}{N}$.

The Normal Distribution

The Mean and the Standard Deviation of a Normal Distribution

Let $\mathcal{D} = \{x_1, \ldots, x_N\}$ and assume that these data are modeled with a normal distribution having mean μ. Then the MLE of the mean $\theta = \mu$ for the normal distribution is:

$$\hat{\theta} = \frac{\sum_{i=1}^{N} x_i}{N}$$

and the MLE of the standard deviation $\theta = \sigma$ is:

$$\hat{\theta} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}.$$

Notice that, in the above estimations, we did not need to know (consider) the standard deviation σ of the normal distribution (since it does not play any role in the derivation details of the above computations).

Reliability of MLE

Merits and Cautions About Maximum Likelihood

- The **consistency** of the MLE is one of its advantages: "consistency" means that the estimation of parameters can be made arbitrarily close to the right value by having a sufficiently large dataset.
- Another important feature of the MLE is that it produces an estimate of parameters that corresponds to the model that is (in some appropriate sense) the closest to the source of the data.
- Nonetheless, there are some serious problems with MLE. A typical problem is in cases when it might be hard to find the maximum of the likelihood exactly (even numerically).
- Another problem with the MLE appears in the case of small data: then the MLE estimates for certain distributions are found to be biased for small sample sizes causing discrepancies in analysis.
- Moreover, with MLE there is no mechanism to incorporate prior beliefs. But this is exactly the situation tackled by the methodology of Bayesian inference, as we are going to examine next.

Bayesian Inference

Theorem (**Bayes' Rule**)

Suppose we have some subjective beliefs or some prior information about parameter(s) θ, before gathering the data $\mathcal{D}$. Moreover, we would like to take this information into account when we estimate our model of the data. One way to do so is to place a **prior probability distribution** $P(\theta)$ on the parameters θ. Then, applying **Bayes' rule**, we get the **posterior probability distribution** $P(\theta|\mathcal{D})$ as:

$$\underbrace{P(\theta|\mathcal{D})}_{\text{Posterior}} = \frac{\overbrace{P(\mathcal{D}|\theta)}^{\text{Likelihood}} \times \overbrace{P(\theta)}^{\text{Prior}}}{\underbrace{P(\mathcal{D})}_{\text{Normalizing constant}}}.$$

Extracting information from the posterior $P(\theta|\mathcal{D})$ is usually called **Bayesian inference**.

Maximum A Posteriori Probability (MAP) Estimation

Procedure (**Maximum A Posteriori Probability (MAP) Estimation**)

The **Maximum A Posteriori Probability** (**MAP**) estimate chooses $\theta = \hat{\theta}$, where the posterior $P(\theta|\mathcal{D})$ is maximized as a function of θ:

$$\hat{\theta} = \arg\max_{\theta} P(\theta|\mathcal{D}).$$

Remarks

- Notice that to get the MAP estimate $\hat{\theta}$, we do not need to know the value of the posterior, since it is enough to work with the product:

$$\begin{aligned} P(\theta|\mathcal{D}) &\propto P(\mathcal{D}|\theta)P(\theta) \\ &\propto \text{Likelihood} \times \text{Prior}. \end{aligned}$$

- Similarly, the value of the normalizing constant $P(\mathcal{D})$ does not matter (because it is a multiplicative constant that cannot change the location of the maximum).
- It can be easily shown that MLE is a special case of MAP, when the prior is uniform.

Conjugacy

Definition (**Conjugate Distributions and Priors**)

If the posterior probability distribution $P(\theta|\mathcal{D})$ is in the same family of probability distributions as the prior probability distribution $P(\theta)$, then the prior and the posterior are called **conjugate** probability distributions. Moreover, the prior is called a **conjugate prior** for the likelihood function $P(\mathcal{D}|\theta)$.

Usefulness of Conjugate Priors

- Conjugate priors are usually useful when we want to derive a closed-form expression for the posterior distribution.
- They can be easily interpreted, since they allow us to know how the parameters of the prior change after the Bayesian update.

Cautions About Bayesian Inference

Possible Problems with the MAP Estimation

- When there is a lot of data, the prior has little influence on the outcome of the inference and the MAP estimation looks a lot like the MLE solution.
- Thus, the two estimations differ substantially, only when there is little data, in which case the prior matters.
- However, the difficulty with the MAP approach is that it might be hard to select a "good" prior.
- Notice that the inconvenience to justify the choice of a prior represents a significant philosophical problem, which examines the fervidly debated and contested relationship between statistics and induction. (A standard reference for the latter discussion is the Stanford Encyclopedia of Philosphy (SEP) article on the "Philosophy of Statistics" written by Jan-Willem Romeijn.)