ECEN 740: Machine Learning Engineering

Lecture 1: Bayesian Point Estimation

Tie Liu

Texas A&M University

Outline

Topics:

- The estimation problem
- The frequentist approach:
  - The maximum-likelihood (ML) estimator
  - The frequentist risk
- The Bayesian approach
  - The maximum-a-posteriori (MAP) estimator
  - The Bayesian risk and Bayes estimator

Main reference:

- *The Bayesian Choice* by Christian P. Robert (`https://link.springer.com/book/10.1007/0-387-71599-1`)

Problem setup

- Let $\theta \in \Theta$ be an *unknown* parameter whose value dictates the distribution of an observation $\mathsf{x} \in \mathcal{X}$. The conditional density $p(x|\theta)$ is known as the *likelihood* and is assumed to be *known* to the decision maker

- The goal of the decision maker is to *predict* the value of θ based on the value of x

- Formally, a *prediction rule*, or an *estimator*, is a function $\delta : \mathcal{X} \to \Theta$

- In statistics, there are two different approaches for constructing a good estimator: The *frequentist* approach and the *Bayesian* approach

The maximum-likelihood (ML) estimator

- In the frequentist approach, the parameter θ is assumed to be *deterministic (but unknown)*, and the goal is to find an estimator that performs well under *repeated* observations (on average or with high probability)

- A prototypical frequentist estimator is the *maximum-likelihood (ML)* estimator:

$$\delta_{ML}(x) := \arg\max_{\theta\in\Theta} p(x|\theta)$$

- While for many models the ML method produces highly intuitive and well-performed estimators, it does *not* always work to full satisfaction – there are models where the ML method gives poor or even pathological estimates

- Next, let us look at two examples: Estimating *Gaussian mean* and estimating *categorical probabilities*

Estimating Gaussian mean

- Let $\mathsf{x} = (\mathsf{x}_1, \ldots, \mathsf{x}_n)$ be a collection of n independent observations such that $\mathsf{x}_i \sim \mathcal{N}(\theta, \sigma^2)$, where the mean θ is the *unknown* parameter and the variance σ^2 is assumed to be *known*
- In this case, the likelihood is given by

$$
\begin{aligned}
p(x|\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i-\theta)^2}{2\sigma^2}\right] \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right]
\end{aligned}
$$

- The ML estimator is thus given by

$$
\delta_{ML}(x) = \arg\max_{\theta} p(x|\theta) = \arg\min_{\theta} \sum_{i=1}^{n}(x_i-\theta)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i
$$

i.e., the *empirical-mean* estimator

Estimating categorical probabilities

- Consider n independent draws from a categorical distribution over $[K]$ with unknown parameters $\theta = (p_1, \ldots, p_K)$. Let $\mathsf{x} = (\mathsf{n}_1, \ldots, \mathsf{n}_K)$ where n_k is the number of occurrences of category k from the n draws

- In this case, the likelihood is given by

$$p(x|\theta) = \frac{n!}{n_1! \cdots n_K!} \prod_{k=1}^{K} p_k^{n_k}$$

- To maximize the likelihood $p(x|\theta)$, let $r = (r_1, \ldots, r_K)$ where

$$r_k = \frac{n_k}{n}$$

  is the *empirical* probability for category k

- We have

$$\begin{aligned}
&-\frac{1}{n}\log\left[\frac{n_1!\cdots n_K!}{n!}p(x|\theta)\right] \\
&\qquad= -\frac{1}{n}\sum_{k=1}^{K} n_k \log p_k = \sum_{k=1}^{K} r_k \log\frac{1}{p_k} \\
&\qquad= \sum_{k=1}^{K} r_k \log\frac{r_k}{p_k} - \sum_{k=1}^{K} r_k \log r_k \\
&\qquad= D_{KL}(r\|\theta) - \sum_{k=1}^{K} r_k \log r_k \\
&\qquad\geq -\sum_{k=1}^{k} r_k \log r_k
\end{aligned}$$

for any probability vector $\theta = (p_1, \ldots, p_K)$. Here, $D_{KL}(r\|\theta)$ is the *Kullback-Leibler (KL) divergence* between the empirical probability vector r and the (true) probability vector θ

- The previous inequality is due to the *non-negativity* of KL divergence, and it holds with an equality if and only if $\theta = r$

- The ML estimator is thus given by

$$\begin{aligned}\delta_{ML}(x) &= \arg\max_{p} p(x|\theta) \\ &= \arg\min_{p} \left\{ -\log \left[\frac{n_1! \cdots n_K!}{n!} p(x|\theta) \right] \right\} \\ &= \left(\frac{n_1}{n}, \ldots, \frac{n_K}{n} \right)\end{aligned}$$

  i.e., the *empirical-probability* estimator

- While in this case the ML method again produces a highly intuitive estimator, the estimated probabilities are *zero* for any *unseen* categories. One can argue that in this case, the ML estimator leans too heavily on empirical evidence and lacks *foresight*

- The pathological nature of the empirical-probability estimator was first recognized in the 18th century by Pierre-Simon Laplace in the context of the *sunrise problem*

The frequentist risk

- To formally evaluate the performance of an estimator requires a *loss function*

- A loss function is any function $\ell : \Theta \times \Theta \to \mathbb{R}_+$ such that $\ell(\theta, \hat{\theta})$ measures the *inaccuracy* of the estimate $\hat{\theta}$ when the unknown parameter is θ

- In the eye of a frequentist, the unknown parameter θ is *fixed* but the observation x can be made *repeatedly*. It is thus natural to evaluate the performance of an estimator δ by the following *frequentist risk*:
$$R_\delta(\theta) := \mathbb{E}_\theta[\ell(\theta, \delta(\mathsf{x}))]$$
where the expectation is over $\mathsf{x} \sim p(x|\theta)$

- Note that the frequentist risk associated with an estimator is not a number but rather a *function* (of the unknown parameter). Since the space of functions is not totally ordered, the concept of the frequentist risk does not lend itself useful in terms of constructing *optimal* estimators

The maximum-a-posteriori (MAP) estimator

- In the Bayesian setting, the unknown parameter θ is assumed to be *random* with a prior density $p(\theta)$, which reflects our *predisposition* on the possible values that the unknown parameter may take

- The existence of a prior density allows us to compute the *posterior* density $p(\theta|x)$ according to the Bayes theorem:

  $$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)}$$

  where the marginal density

  $$p(x) = \int p(\theta)p(x|\theta)d\theta$$

  is also known as the *evidence* of x. The product $p(\theta)p(x|\theta)$ is sometimes referred to as the *un-normalized posterior*

- Note that the posterior density is a conditional density conditioned on the value of a *single* observation, so this quantity is completely meaningless to a frequentist

- Bayesian statisticians, on the other hand, are perfectly fine with the concept of posterior density, and the following *maximum-a-posteriori (MAP)* estimator is completely natural to them:

$$\delta_{MAP}(x) := \arg\max_{\theta \in \Theta} p(\theta|x)$$

- By the Bayes theorem, the MAP estimator can be equivalently written as

$$\delta_{MAP}(x) = \arg\max_{\theta \in \Theta} \{p(\theta)p(x|\theta)\}$$

  or

$$\delta_{MAP}(x) = \arg\max_{\theta \in \Theta} \{\log p(\theta) + \log p(x|\theta)\}$$

  where these last two expressions have the advantage of not needing to evaluate the evidence integral

- Finally, note that if $p(\theta)$ is *constant* over Θ (a *uniform* prior), the MAP estimator reduces to the ML estimator (even though in the eyes of a frequentist and a Bayesian statistician, they come from very different perspectives)

The Bayesian risk

- Recall that in the frequentist approach, the unknown parameter θ is *deterministic* and the performance of an estimator can be formally evaluated using the *frequentist* risk:

  $$R_\delta(\theta) = \mathbb{E}_\theta[\ell(\theta, \delta(\mathsf{x}))]$$

  where the expectation is over $\mathsf{x} \sim p(x|\theta)$

- In the Bayesian approach, the unknown parameter θ is *random* and the performance of an estimator can be formally evaluated using the *Bayesian risk*

  $$R_\delta := \mathbb{E}[\ell(\theta, \delta(\mathsf{x}))]$$

  where the expectation is over the *joint* distribution of (θ, x)

- By the law of total expectation, the Bayesian risk

  $$R_\delta = \mathbb{E}[\mathbb{E}[\ell(\theta, \delta(\mathsf{x}))|\theta]] = \mathbb{E}[R_\delta(\theta)]$$

  where the last expectation is over the prior distribution of θ

Bayes estimator

- Note that unlike the frequentist risk (which is a function), the Bayesian risk is a number. Therefore, not only can it be used to evaluate the performance of an estimator, it can also be used to *construct* good estimators

- More specifically, we say an estimator δ is a *Bayes estimator* if it minimizes (among all possible estimators) the Bayesian risk for some loss function ℓ

- For a given loss function ℓ, to find the corresponding Bayes estimator, let us define the *posterior risk*

  $$R_\delta(x) := \mathbb{E}_x\left[r(\theta, \delta(x))\right]$$

  where the expectation is over the *posterior* distribution $p(\theta|x)$ of θ

- Again by the law of total expectation, we can write the Bayesian risk R_δ as

$$R_\delta = \mathbb{E}\left[\mathbb{E}\left[r(\theta, \delta(\mathsf{x}))|\mathsf{x}\right]\right] = \mathbb{E}[R_\delta(\mathsf{x})]$$

  where the last expectation is over the marginal distribution of x

- Note that the posterior risk $R_\delta(x)$ only depends on the value of δ at x. Since we have *no* restrictions on δ, to minimize the Bayesian risk R_δ, we can choose δ to minimize the posterior risk $R_\delta(x)$ *separately* at each x:

$$\delta_B(x) = \arg\min_{\hat{\theta}\in\Theta} R_{\hat{\theta}}(x)$$

  where

$$R_{\hat{\theta}}(x) := \mathbb{E}_x[r(\theta, \hat{\theta})]$$

  is the posterior risk that corresponds to the estimate $\delta(x) = \hat{\theta}$

Classification under the 0-1 loss

- Recall that when the parameter θ is *categorical*, an estimation problem is also known as a *classification* problem
- For classification problems, a commonly used loss function is the *0-1 loss*:

  $$\ell_{0-1}(\theta, \hat{\theta}) := 1_{\{\theta \neq \hat{\theta}\}}$$

  Note that the 0-1 loss only penalizes classification errors but does *not* discriminate against any particular error patterns
- Under the 0-1 loss, the posterior risk

  $$R_{\hat{\theta}}(x) = \sum_{\theta \in \Theta} 1_{\{\theta \neq \hat{\theta}\}} p(\theta|x) = \sum_{\theta \neq \hat{\theta}} p(\theta|x) = 1 - p(\hat{\theta}|x)$$

- The Bayes estimator is thus given by

  $$\delta_B(x) = \arg\min_{\hat{\theta} \in \Theta} \left\{1 - p(\hat{\theta}|x)\right\} = \arg\max_{\hat{\theta} \in \Theta} p(\hat{\theta}|x)$$

  i.e., it reduces to the MAP estimator. In another word, for *classification* problems, the MAP estimator is the Bayes estimator under the 0-1 loss

Regression under the squared-error loss

- Recall that when the parameter θ is *numerical*, an estimation problem is also known as a *regression* problem

- For regression problems, a commonly used loss function is the *squared-error loss*:
$$\ell_{se}(\theta, \hat{\theta}) := (\theta - \hat{\theta})^2$$

- Under the squared-error loss, the posterior risk
$$\begin{aligned} R_{\hat{\theta}}(x) &= \mathbb{E}_x[(\theta - \hat{\theta})^2] \\ &= \mathbb{E}_x[(\theta - \mathbb{E}_x[\theta] + \mathbb{E}_x[\theta] - \hat{\theta})^2] \\ &= \mathbb{E}_x[(\theta - \mathbb{E}_x[\theta])^2] + (\mathbb{E}_x[\theta] - \hat{\theta})^2 \end{aligned}$$

- The Bayes estimator is thus given by
$$\delta_B(x) = \mathbb{E}_x[\theta] = \int_\Theta \theta p(\theta|x) d\theta$$

  That is, under the squared-error loss, the Bayes estimator is the *posterior-mean estimator*

Bayesian inference

- Based on our previous discussions, it is clear that the posterior density $p(\theta|x)$ plays a *central* in Bayesian statistics

- Note that while some of the Bayes estimators can be computed from the *un-normalized* posterior density $p(\theta)p(x|\theta)$, the others such as the conditional mean estimator can only be computed from the (normalized) posterior density $p(\theta|x)$

- In Bayesian statistics, the task of computing the posterior density $p(\theta|x)$ from the prior $p(\theta)$ and the likelihood $p(x|\theta)$ is known as *inference*, for which the main challenge is that the *evidence integral*

  $$p(x) = \int_{\Theta} p(\theta)p(x|\theta)d\theta$$

  can be very difficult to evaluate

- An useful concept that can lead to a *closed-form* expression for the evidence integral is *conjugate prior*

Conjugate prior

- A *conjugate* prior is defined as a prior distribution belonging to some parametric family, for which the resulting posterior distribution also belongs to the *same* family

- Conjugate priors are especially useful for *sequential* estimation, where the posterior of the current measurement is used as the prior in the next measurement

- As a simple example, consider the estimation problem for which the likelihood $\mathsf{x}|\theta \sim \mathcal{N}(\theta, \sigma^2)$. In this case, if the prior is chosen as $\theta \sim \mathcal{N}(\mu, \tau^2)$, the posterior is also Gaussian and the posterior-mean estimator is given by

$$\delta_B(x) = \frac{\tau^2}{\sigma^2 + \tau^2} x + \frac{\sigma^2}{\sigma^2 + \tau^2} \mu$$

- For a more involved example, next let us return to the problem of estimating categorical probabilities that we discussed earlier

Estimating categorical probabilities (cont'd)

- Recall that in this problem, the unknown parameter is the probability vector $\theta = (p_1, \ldots, p_K)$ and the likelihood

  $$p(x|\theta) = \frac{n!}{n_1! \cdots n_K!} \prod_{k=1}^{K} p_k^{n_k}$$

  is a *multinomial* distribution

- Assume that θ is *uniform* over the $(K-1)$-simplex

- To calculate the posterior density $p(\theta|x)$, we use the facts that i) a uniform distribution over a simplex is a member of the *Dirichlet* distribution family; and ii) Dirichlet distributions are *conjugate* priors to a multinomial likelihood

- It follows that the *posterior-mean* estimator is given by

  $$\delta_B(x) = \left(\frac{n_1 + 1}{n + K}, \ldots, \frac{n_K + 1}{n + K} \right)$$

- In literature, this is known as *Laplace's add-one estimator* or *Laplace's rule of succession*

- Compared with the ML estimator

  $$\delta_{ML}(x) = \left(\frac{n_1}{n}, \ldots, \frac{n_K}{n}\right)$$

  the add-one estimator will assign a small but *nonzero* probability to categories with zero occurrences, a feature that can be very important to some applications (as we shall see shortly)

- Laplace's add-one estimator is perhaps the simplest *smoothing* technique for estimating the probability of *unseen* events. A more sophisticated smoothing technique known as the *Good-Turing estimator* played a role in breaking the Enigma machine during World War II

Approximate inference

- While the concept of conjugate prior is extremely useful for computing the posterior from the prior and the likelihood, for many models the evidence integral is unavailable in closed form or requires exponential time to compute - this is why exact inference in such models is hard and *approximate* inference is needed
- In the next lecture, we shall discuss an increasingly popular approximate inference technique known as *variation inference*