

DS 102 Discussion 3

Wednesday, September 15, 2021

1. Decision Theory: Computing and Minimizing the Bayes Risk

For the following two parts, derive the decision procedure δ^* that minimizes the Bayes risk, for the given loss function. That is, provide an expression for

$$\delta^* = \underset{\delta}{\operatorname{argmin}} R(\delta)$$

where the Bayes risk $R(\delta)$ can be written out as

$$R(\delta) = \mathbb{E}_{\theta, X}[\ell(\theta, \delta(X))] = \mathbb{E}_X[\mathbb{E}_{\theta}[\ell(\theta, \delta(X)) \mid X]].$$

Hint. One strategy to find the decision rule that minimizes the Bayes risk is based on the following rationale. For any given value of the data, $X = x$, the quantity $\delta(x)$ is simply a scalar value. Suppose, for any given value of $X = x$, we can find the scalar value $\delta^*(x) = a^* \in \mathbb{R}$ such that

$$a^* = \underset{a \in \mathbb{R}}{\operatorname{argmin}} \mathbb{E}_{\theta}[\ell(\theta, a) \mid X = x]$$

(that is, a^* is the scalar value that minimizes the Bayes posterior risk for this particular value of $X = x$). Then, the rule given by this computation of $\delta^*(x)$ (for each value of $X = x$) must also be the one that minimizes the Bayes risk, which just takes an expectation over all possible values of X . This is sometimes referred to as a *pointwise minimization* strategy.

(a) $\ell(\theta, \delta(X)) = (1/2)(\theta - \delta(X))^2$ (squared-error loss)

(b) $\ell(\theta, \delta(X)) = \mathbf{1}[\theta \neq \delta(X)]$ (zero-one loss)



2. Conjugate Priors

In this question, we will investigate examples of *conjugate priors*: pairs of distributions (for the likelihood and the prior) such that the prior and posterior are from the same distribution, with possibly different parameters.

Recall that for observed data X , and prior distribution $p(\theta)$ on parameters θ , the *posterior probability* distribution on θ , after seeing the data, is given by¹

$$\begin{aligned} p(\theta|x) &= \frac{p(x|\theta) \cdot p(\theta)}{p(x)} \\ &\propto p(x|\theta) \cdot p(\theta) \end{aligned}$$

where \propto denotes “proportional to.” Note here that $p(x)$ is a normalization constant which allows the posterior distribution to sum to 1. However, it bears no influence on the shape of the posterior distribution because it doesn’t contain θ . Therefore, we can always work this proportionality to try to identify a posterior distribution.

(a) Beta and Binomial

Say you’ve observed a sequence of coin flips, X_1, \dots, X_n , all using the same coin, which has some probability of landing heads, p_h . Denote by H the total number of heads:

$$H = \sum_{i=1}^n \mathbb{I}\{X_i = \text{heads}\}$$

H follows a binomial distribution, with PDF

$$p(H = k) = \binom{n}{k} p_h^k (1 - p_h)^{n-k}$$

We didn’t make this coin, it was given to us. We’re willing to place a prior distribution on the probability of it landing heads and we’ll use the beta distribution to do so. The beta distribution is a suitable choice since it takes on values from $[0,1]$, which can be used to model probabilities. The beta distribution PDF is parameterized by shape parameters $\alpha > 0$ and $\beta > 0$, and is given by

$$f(z; \alpha, \beta) = \frac{(\alpha + \beta - 1)!}{(\alpha - 1)! (\beta - 1)!} z^{\alpha-1} (1 - z)^{\beta-1}, \quad 0 < z < 1$$

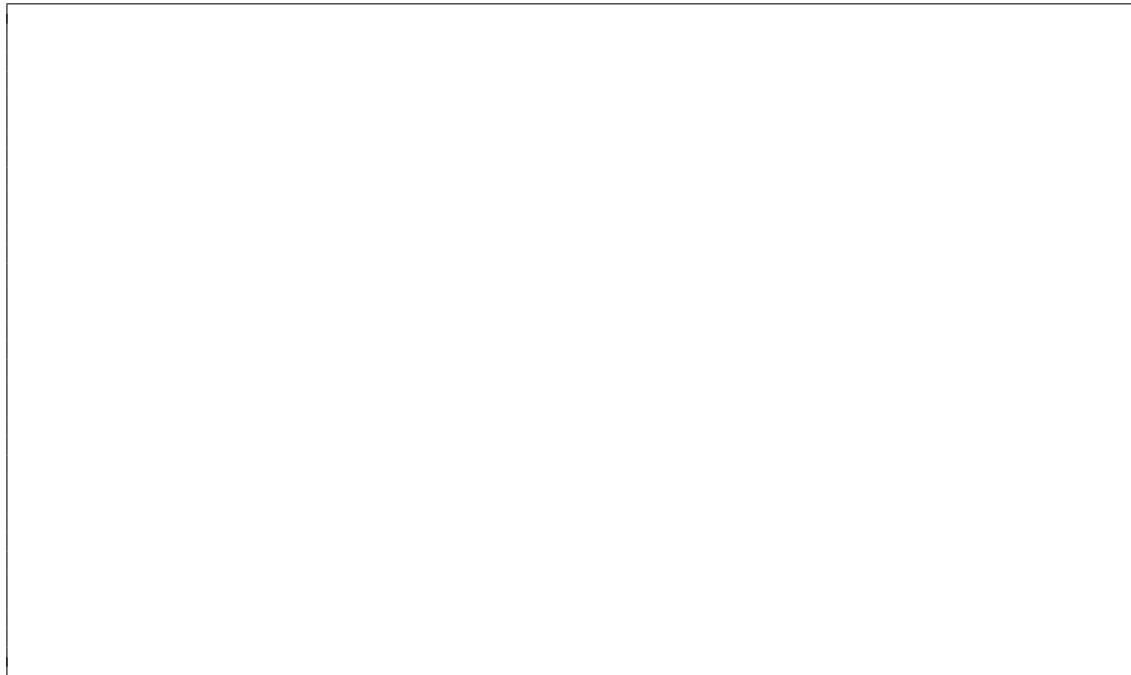
Show that the beta distribution is a conjugate prior for the binomial distribution. What are the shape parameters for the posterior distribution?

¹The *prior* distribution on the parameters is given by $p(\theta)$ and the likelihood $p(x|\theta)$.



(b) *Gamma and Exponential*

A gamma distribution with parameters α, β has density function $p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ where $\Gamma(\alpha)$ is the gamma function (see https://en.wikipedia.org/wiki/Gamma_distribution). Show that gamma distribution is a conjugate prior for exponential distribution for multiple measurements, i.e. if we have samples X_1, X_2, \dots, X_n that are mutually independent given λ , and each $X_i | \lambda \sim \text{Exp}(\lambda)$ and $\lambda \sim \text{Gamma}(\alpha, \beta)$, then $\lambda | X_1, X_2, \dots, X_n \sim \text{Gamma}(\alpha^*, \beta^*)$ for some values α^*, β^* .



3. Parameter Estimation: MLE vs. MAP

In this question, we will review two parameter estimation strategies called *Maximum Likelihood Estimation* (MLE) and *Maximum a Posteriori* (MAP) Estimation. Both strategies aim to provide an estimate for the value of a parameter of a distribution θ , based on some data collected X .

Assuming we know the type of distribution from which our data X was drawn from, we can estimate the distribution's parameter θ with MLE in the following way:

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}} p(X|\theta)$$

In other words, MLE finds the most likely value of the fixed parameter θ , given the data. Similarly, the MAP Estimate also takes into the account the likelihood of the data, given the parameter θ . However, the MAP Estimate also incorporates a prior probability of θ . It is given by:

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}} p(X|\theta)p(\theta)$$

Therefore, the MAP Estimate finds the value of the random parameter θ which is most probable, given the data and a prior belief.

(a) MLE for Binomial Distribution


Recall that the PMF of a Binomial random variable X is given by

$$P(X = k; p_H) = \binom{n}{k} p_H^k (1 - p_H)^{n-k}$$

Find the MLE for p_H , the chance of success.

(b) MAP for Binomial Distribution, with Beta Prior

Find the MAP Estimate for p , the chance of success. Compare your result to the MLE found in Part (a).



(c) Connecting MAP and MLE

Compare the estimates of p in the Parts (a) and (b). What is the relationship between the MLE and MAP Estimate of a parameter θ ?

