DS 102 Discussion 3

Wednesday, February 9, 2022

1. **Decision Theory: Computing and Minimizing the Bayes Risk**

For the following two parts, derive the decision procedure δ^* that minimizes the Bayes risk (not the same as the Bayesian posterior risk), for the given loss function. That is, provide an expression for

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

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

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

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

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

(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) (Optional) $\ell(\theta, \delta(X)) = \mathbf{1}[\theta \neq \delta(X)]$ (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 by[1]

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

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, $X_1, ..., X_n$, 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 parameters of the new Beta distribution?

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

(b) (Optional) *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 $X_1, X_2, \cdots, X_n$ that are mutually independent given λ, and each $X_i|\lambda \sim Exp(\lambda)$ and $\lambda \sim Gamma(\alpha, \beta)$, then $\lambda|X_1, X_2, \cdots, X_n \sim Gamma(\alpha^*, \beta^*)$ 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_h, the chance of success. Compare your result to the MLE found in Part (a).

Hint 1: Use the result from 2(a).

Hint 2: The mode of a Beta(α, β) distribution is $\frac{\alpha-1}{\alpha+\beta-2}$.

(c) Connecting MAP and MLE

Compare the estimates of p in the Parts (a) and (b). When would the MLE and MAP Estimates for θ be equal to each other?

4. **Graphical Models**

Last lecture, we were introduced to *Graphical Models*, which are flexible diagrams to express the relationships between random variables. An important special case of graphical models are Bayesian hierarchical models, which generally may look like the figure below:

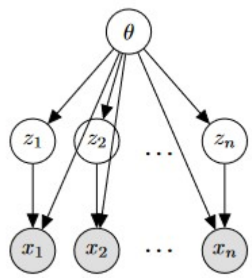


Figure 1: **Bayesian hierarchical model with hyperparameter θ, latent variables z_i, and observed variables x_i**

In a Bayesian hierarchical model, observations are independent given the latent variables, and each observed variable depends only on its corresponding latent variable and the hyperparameters. As a result, Bayesian hierarchical models are always depicted as *directed acyclic graphs* (DAGs).

In the following subparts, we will create our own graphical model and explore its properties.

(a) *Formulating a Graphical Model*

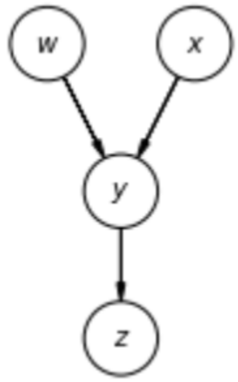
Suppose you are a farmer who wants to model the upcoming crop harvest. You are interested in the following variables:

- w is the amount of pesticide used
- x is the amount of total rainfall for the season
- y is the number of bugs found in the field
- z is the total crop yield

Draw a graphical model to illustrate the relationships between these variables.

(b) *Identifying Independence and Conditional Independence*

Consider the following graphical model:



Which of the following statements are true about the graphical model above?

1. $x \perp\!\!\!\perp w$
2. $w \perp\!\!\!\perp x \mid y$
3. $w \perp\!\!\!\perp z \mid y$