Introduction to Bayesian Decision Theory

Angela J. Yu

1 Introduction

In the Bayesian framework, we assume that observable data $\mathbf{x}$ are generated by underlying hidden *causes* $\mathbf{s}$ in the world, which cannot be observed directly. The *generative model* specifies how the data gets generated from the causes, which is encapsulated in the conditional probability $p(\mathbf{x}|\mathbf{s})$, and any prior information about the distribution over the different states of the causes $p(\mathbf{s})$. The generative probability distributions may be parameterized by certain model parameters, so we sometimes write $p(\mathbf{x}|\mathbf{s};\boldsymbol{\theta})$ and $p(\mathbf{s};\boldsymbol{\phi})$, where $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ parameterize their respective distributions.

The *recognition model* involves computing the posterior distribution $p(\mathbf{s}|\mathbf{x})$, which is given by Bayes' Theorem:

$$\begin{aligned} p(\mathbf{s}|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathbf{s})p(\mathbf{s})}{p(\mathbf{x})} \\ &= \frac{p(\mathbf{x}|\mathbf{s})p(\mathbf{s})}{\int p(\mathbf{x}|\mathbf{s}')p(\mathbf{s}')d\mathbf{s}'} \\ &= \frac{p(\mathbf{x}|\mathbf{s})p(\mathbf{s})}{Z} \\ &\propto p(\mathbf{x}|\mathbf{s})p(bs) \end{aligned}$$

where $Z = p(\mathbf{x})$ is the normalization constant independent of $\mathbf{s}$ (due to the integration). Implicit in this formulation is the dependence on the model parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, so that $p(\mathbf{x}|\mathbf{s}) = \int p(\mathbf{x}|\mathbf{s};\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ and $p(\mathbf{s}) = \int p(\mathbf{s};\boldsymbol{\phi})p(\boldsymbol{\phi})d\boldsymbol{\phi}$. Computing the distributions $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\phi})$, typically based on previously observed data $\mathcal{D}_n \equiv \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, constitute the *learning* problem. In theory, this can be achieved again using Bayes' Theorem; in practice, exact learning is often intractable, and so point estimates for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are used, or samples are drawn from the posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathcal{D}_n)$ to approximate the integration.

Given the observation $\mathbf{x}$, it is often necessary to report or act on an estimate of $\mathbf{s}$ that "best" describes the data. Bayesian decision theory formalizes this process of translating information into action. First, we define the loss function $l_{\mathbf{x}}(\hat{\mathbf{s}}, \mathbf{s}^*)$, which quantifies the *loss* or *cost* associating with report $\mathbf{s} = \hat{\mathbf{s}}$ when the data were actually generated by $\mathbf{s} = \mathbf{s}^*$. Then we can compute the *average* or *expected loss*:

$$\mathcal{L}_{\mathbf{x}}(\hat{\mathbf{s}}) = \langle l_{\mathbf{x}}(\hat{\mathbf{s}}, \mathbf{s}^*) \rangle_{p(\mathbf{s}=\mathbf{s}^*|\mathbf{x})} = \int l_{\mathbf{x}}(\hat{\mathbf{s}}, \mathbf{s}^*) p(\mathbf{s} = \mathbf{s}^*|\mathbf{x}) d\mathbf{s}^* \tag{1}$$

The expected loss is integrated over all possible settings of $\mathbf{s}$, weighed by their relative probabilities, and indicates how much loss can be expected when $\hat{\mathbf{s}}$ is chosen as the estimate. The optimal decision procedure has to choose a $\hat{\mathbf{s}}$ that minimizes this expected loss. In the following, we look at three concrete and common examples of loss functions and how this framework can be applied in each case. For simplicity, we will assume in all these examples that the hidden variable is scalar and denote it as s.

1.1 Binary (0-1) Loss

The 0-1 binary loss function has the following form:

$$l_{\mathbf{x}}(\hat{s}, s^*) = 1 - \delta_{\hat{s}s^*} = \begin{cases} 1 & \text{if } s^* \neq \hat{s} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where δ is the Kronecker Delta function. This loss function is an appropriate choice when the same penalty is incurred whenever the estimate does not exactly correspond to the true underlying variable (and none otherwise), regardless of how far from the truth the estimate actually is. It makes most sense when the hypothesis space (the space of values s can take on) is discrete. Substituting Eq. 2 into Eq. 1, the expected loss is:

$$\begin{aligned} \mathcal{L}_{\mathbf{x}}(\hat{s}) &= \sum_{s^*} l_{\mathbf{x}}(\hat{s}, s^*) P(s = s^*|\mathbf{x}) \\ &= \sum_{s^*} (1 - \delta_{\hat{s}s^*}) P(s = s^*|\mathbf{x}) \\ &= \sum_{s^*} P(s = s^*|\mathbf{x}) ds^* - \sum_{s^*} \delta_{\hat{s}s^*} P(s = s^*|\mathbf{x}) \\ &= 1 - P(s = \hat{s}|\mathbf{x}) \end{aligned}$$

Clearly, this quantity is minimized when $\hat{s}$ is chosen to be the maximum of the posterior distribution $P(s|\mathbf{x})$, or the *max a posteriori* (MAP) estimate. The corresponding minimal loss that can be achieved is $1 - P(s^*|\mathbf{x})$.

1.2 Square Loss

The square loss function is defined as follows:

$$l_{\mathbf{x}}(\hat{s}, s^*) = (\hat{s} - s^*)^2 \tag{3}$$

Unlike the binary loss function, the square loss function does care about *how* different the estimated $\hat{s}$ is from the true s^*. It is most appropriate when s lives in a continuous space with a well-defined metric, so that it makes sense to look at the square distance two different values of s. Substituting Eq. 3 into Eq. 1, we get the following:

$$\begin{aligned} \mathcal{L}_{\mathbf{x}}(\hat{s}) &= \int l_{\mathbf{x}}(\hat{s}, s^*) p(s = s^*|\mathbf{x}) ds^* \\ &= \int (\hat{s} - s^*)^2 p(s = s^*|\mathbf{x}) ds^* \\ &= \hat{s}^2 \int p(s = s^*|\mathbf{x}) ds^* - 2\hat{s} \int s^* p(s = s^*|\mathbf{x}) ds^* + \int s^{*2} p(s = s^*|\mathbf{x}) ds^* \\ &= \hat{s}^2 - 2\hat{s}\langle s^* \rangle + \langle s^{*2} \rangle \\ &= \hat{s}^2 - 2\hat{s}\langle s^* \rangle + \langle s^* \rangle^2 + \langle (s^* - \langle s^* \rangle)^2 \rangle \\ &= (\hat{s} - \langle s^* \rangle)^2 + \langle (s^* - \langle s^* \rangle)^2 \rangle \end{aligned}$$

where all the expectations are taken over $p(s = s^*|\mathbf{x})$. Note that the second term in the last line is independent of $\hat{s}$, and the first term is minimized when $\hat{s} = \langle s^* \rangle$. Thus, the expected square loss is minimized when $\hat{s}$ is chosen to be the expectation of s under the posterior distribution, and the minimal loss is the covariance of this distribution.

1.3 Absolute Loss

Another popular loss function that takes into account the quality of the estimate is the absolute function:

$$l_{\mathbf{x}}(\hat{s}, s^*) = |\hat{s} - s^*|. \tag{4}$$

At first glance, this loss function may seem very similar to the square loss function. But we shall soon see that minimizing this loss leads to a very different answer than the square loss case. Substituting Eq. 4 into Eq. 1, we have the following:

$$\begin{aligned}
\mathcal{L}_{\mathbf{x}}(\hat{s}) &= \int l_{\mathbf{x}}(\hat{s}, s^*) p(s{=}s^*|\mathbf{x}) ds^* \\
&= \int |\hat{s} - s^*| p(s^*|\mathbf{x}) ds^* \\
&= \int_{-\infty}^{\hat{s}} (\hat{s} - s^*) p(s^*|\mathbf{x}) ds^* + \int_{\hat{s}}^{\infty} (s^* - \hat{s}) p(s^*|\mathbf{x}) ds^* \\
&= \hat{s} \left(\int_{-\infty}^{\hat{s}} p(s^*|\mathbf{x}) ds^* - \int_{\hat{s}}^{\infty} p(s^*|\mathbf{x}) ds^* \right) + \int_{\hat{s}}^{\infty} s^* p(s^*|\mathbf{x}) ds^* - \int_{-\infty}^{\hat{s}} s^* p(s^*|\mathbf{x}) ds^*
\end{aligned}$$

To obtain the optimal value of $\hat{s}$, we set the derivative of $\mathcal{L}_{\mathbf{x}}(\hat{s})$ to 0 and solve for $\hat{s}$. Utilizing the Product Rule and the Fundamental Theorem of Calculus, we obtain the following:

$$\begin{aligned}
\frac{d\mathcal{L}_{\mathbf{x}}(\hat{s})}{d\hat{s}} &= 2\hat{s}p(\hat{s}|\mathbf{x}) + \int_{-\infty}^{\hat{s}} p(s^*|\mathbf{x}) ds^* - \int_{\hat{s}}^{\infty} p(s^*|\mathbf{x}) ds^* - 2\hat{s}p(\hat{s}|bx) && (5) \\
&= \int_{-\infty}^{\hat{s}} p(s^*|\mathbf{x}) ds^* - \int_{\hat{s}}^{\infty} p(s^*|\mathbf{x}) ds^* = 0 && (6)
\end{aligned}$$

This implies that the choice of $\hat{s}$ is optimal when it is the *median* of the posterior distribution. That is, when $P(s < \hat{s}|\mathbf{x}) = P(s > \hat{s}|\mathbf{x})$. Returning to Eq. 4, we see that the minimal expected loss for this loss function is $\int_{\hat{s}}^{\infty} s^* p(s^*|\mathbf{x}) ds^* - \int_{-\infty}^{\hat{s}} s^* p(s^*|\mathbf{x}) ds^*$, which is smaller when the posterior distribution itself is peakier (around the median).

2 Discussion

In this short tutorial, we have defined the basic problem and method of Bayesian decision making, and applied the methodology to three common loss functions. In Bayesian decision making, a well-defined *loss function*, indicating the potential loss incurred by each plausible cause-outcome pairing, is critical. Once this loss function is specified, finding the optimal estimate consists of minimizing the *expected* loss, where the expectation is taken over the posterior distribution over the variable of interest, taking into account any uncertainty over the setting of the variable. Specifically, for the case of binary loss, we showed that the optimal estimate is the MAP estimate (or mode), and the minimal expected loss is the probability that the MAP estimate is incorrect. For the case of square loss, the optimal estimate is the expectation of the variable under the posterior distribution, and the minimal expected loss the covariance of that distribution. For the case of absolute loss, which may on the surface greatly resemble the square loss, we saw that the optimal estimate is the median of the posterior distribution, while the minimal expected loss is $\int_{\hat{s}}^{\infty} s^* p(s^*|\mathbf{x}) ds^* - \int_{-\infty}^{\hat{s}} s^* p(s^*|\mathbf{x}) ds^*$, which is smaller when the posterior distribution is peakier (around the median). Note that in the case of a simple Gaussian distribution, the mode, the mean, and the median are all equivalent quantities. Thus, the details of the loss function is most relevant when the posterior distribution is skewed, multi-modal, or takes on other complex properties.