Date: Thursday, February 21, 2019

Chapter 5: Loss functions in Bayesian Inference

1) What is loss and what are common forms?
2) Aim of Bayesian inference: minimized expected loss
3) Examples

Generally speaking, loss quantifies the difference between predicted (or estimated) values and the truth.

↳ "predicted" means we extrapolate from our model to forecast values outside the training set (ML)

↳ "estimated" we're wondering about differences between our model parameter estimates and the truth within the training set (stats).

With estimation, we are concerned with understanding the mechanism that gives rise to the data that we observe.

Examples:
- Relationship of gene expression and cancer subtypes.
- Relationship of demographics and political party affiliation.
- Relationship between regional brain function and clinical outcomes/diagnoses.

Goal: We'd like to be as "close" to quantifying the true relationship as possible.

↓

reaching unbiased estimators for which we can make inference.
(Think confidence intervals and hypothesis testing).

With **Prediction**, our goal is to train a "reasonable" model on the training set and have the highest accuracy / lowest loss on the test set.

Key aim: Maximize Predictive ability.
↳ Lots of metrics (AUC, PROC, F1 Score, RMSE, etc).

Key difference from estimation:
- We allow some bias in our model parameters to dramatically reduce the variance of the model on new data.
  ↳ This is the problem of the bias-variance trade-off.
- In allowing for bias in the model, we typically lose inferential capabilities because we cannot assess the extent of bias on each parameter.

Example: Ordinary Least Squares (OLS) vs Lasso in Regression

- OLS estimates are unbiased and asymptotically normal.
  ↳ we can do inference (☺ for estimation and stats)
- LASSO estimates are biased (they are being squished to 0) and inference is no longer feasible. However, Lasso models typically outperforms OLS models in prediction.

Typically assess the usefulness of the model using a **loss function**. This can be applied to both training and test set performance. Our aim is to minimize loss.

training ↓ estimation

test set ↓ prediction

In general,

$$L(\theta, \hat{\theta}) = f(\theta, \hat{\theta})$$

↳ Loss is a function of the parameter(s) and the estimated value(s).

Examples: 1) squared-loss:

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$$

2) Asymmetric squared-loss:

$$L(\theta, \hat{\theta}) = \begin{cases} (\theta - \hat{\theta})^2, & \hat{\theta} < \theta \\ c(\theta - \hat{\theta})^2, & \hat{\theta} \geq \theta \end{cases}, \quad 0 < c < 1$$

↳ Pushes $\hat{\theta}$ to be $\geq \theta$

3) Absolute loss:

$$L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$$

↳ more robust to outliers in data.

4) zero-one loss:

$$L(\theta, \hat{\theta}) = \begin{cases} 1, & \theta \neq \hat{\theta} \\ 0, & \theta = \hat{\theta} \end{cases}$$

↳ used often in binary classification

5) Log-loss:

$$L(\theta, \hat{\theta}) = -\theta \log(\hat{\theta}) - (1-\theta)\log(1-\hat{\theta}); \quad \theta \in \{0,1\}, \ \hat{\theta} \in [0,1]$$

Note: - We'd like loss functions to be large wherever $\hat{\theta}$ is "far" from θ.

- we can manipulate weights on different scenarios (see asymmetric squared loss) to promote certain values of $\hat{\theta}$.

To-do: 1) Think through why log-loss makes sense.

2) Generalize zero-one loss to maximize true positives.

↳ $L(\theta, \hat{\theta}) = \begin{cases} 0 & , \hat{\theta} = \theta \\ 1 & , \hat{\theta} = 0, \theta = 1 \\ c & , \hat{\theta} = 1, \theta = 0, \\ & 0 < c < 1. \end{cases}$ → We don't want this, heaviest loss

RAIN EXAMPLE

For all examples considered so far, we compare one estimated value, $\hat{\theta}$, with a single parameter value, θ.

Bayesian Aim: minimize expected loss

- In a Bayesian model, θ has a distribution! We make inference based on its posterior given data X.
- So $L(\theta, \hat{\theta})$ is also a random variable that has a distribution!
- Thus, in Bayesian modeling, our goal is to choose $\hat{\theta}$ that minimizes the expected loss:

$$\mathbb{E}_\theta [L(\theta, \hat{\theta})]$$

$$\parallel$$

aka the risk of $\hat{\theta}$, which the book refers to as $\ell(\hat{\theta})$.

- Well, $\mathbb{E}_\theta [L(\theta, \hat{\theta})]$ is clearly not easy to write down (for example, take an integral over θ of some of the previously defined loss functions)!

- Thankfully, we can use Monte Carlo to approximate! (Yay!)

PSEUDO-CODE: for $i = 1:N$

sample $\theta_i \sim \mathbb{P}(\theta | y)$

Approximate: $\hat{\ell}(\hat{\theta}) = \frac{1}{N} \sum_{i=1}^{N} L(\theta_i, \hat{\theta})$

Key point: Bayesian inference takes into account the variability of the loss function over θ; whereas, frequentist analysis only looks at one value of the loss.