STATS 225: Bayesian Analysis
Elements of Bayesian Inference

Babak Shahbaba

Department of Statistics, UCI

Winter, 2015

Decision theory

- In the Bayesian paradigm, estimation, hypothesis testing, and model selection are special cases of decision problems.
- Decision theory provides a mathematical framework for making decision under uncertainty; that is, when the outcome of an event is not known.
- However, we assume that we know our loss (or gain) when one of the possible outcomes occur.

Decision theory

- We use $\mathcal{V}$ to denote the set of all possible values, v, for unknown variables. We refer to $\mathcal{V}$ as the *outcome space*.
- v could be the value future observations. For example, $\mathcal{V} = \{Head, Tail\}$ when you are tossing a coin.
- Or, it could be the value of a parameter in a model. For example $\mathcal{V} = \mathcal{R}$, when we want to estimate μ, the mean of a normal distribution.
- We present the set of all possible actions, a, as $\mathcal{A}$. We refer to $\mathcal{A}$ as the *action space*.
- If we are predicting the outcome of the next coin toss, $\mathcal{A} = \{Head, Tail\}$.
- If we want to estimate μ (i.e., *point estimation*), our action space would be $\mathcal{A} = \mathcal{R}$.
- For hypothesis testing, we can define our action $\mathcal{A} = \{0, 1\}$, where 0 means do not reject the null hypothesis $H_0 : \mu \leq 0$ and 1 means rejecting it.

Decision theory

- We define *Utility* as a function $u = U(v, a)$ that maps the product of outcome space and action space to a real number $u \in \mathcal{R}$ representing how much we gain if we choose action a and the outcome v occurs.
- It is more common to choose a loss function instead of utility (e.g., negative of utility) representing our loss when we choose action a and the outcome v occurs.
- In the coin tossing experiment, the loss function, $L(v, a)$ can be defined as follows:
- $L(Head, Head) = L(Tail, Tail) = 0, L(Head, Tail) = L(Tail, Head) = 1.$
- This is known as $0 - 1$ loss function.

Decision rule

- Now, assume that we have observed data y, for example, $y = HHTHTHHT$, which is the outcome from a sequence of coin tossing. Using this data, we want to make a decision about what the outcome of the next toss would be (or what is θ, the probability of head for this coin).

- The tool for making decision is called *decision rule*, and it's denoted as $\delta(y)$. Note that δ is function of data (i.e., y) only.

- For example, we might define our decision rule for guessing what would be the outcome of the next toss as follows:

$$\delta(y) = \begin{cases} \textit{Head} & \text{if the observed fraction of Heads is } \geq 0.5 \\ \textit{Tail} & \text{if the observe fraction of Heads is } < 0.5 \end{cases}$$

Posterior risk

- Posterior risk for a decision rule is

$$r(\delta|y) = \int_{\mathcal{V}} L(v, \delta(y)) P(v|y) dv$$

- Note that we replaced the action a with the decision rule $\delta(y)$ since our action now depends on our decision rule which itself depends on the observed data.

- Also, note that $p(v|y)$ is the posterior predictive probability if v is future observation (i.e., what is the outcome of the next toss), or it is posterior probability if v is the parameter of a model (i.e., μ, the mean of a normal distribution).

Formal Bayes rule

- *The expected loss* principle: In deciding between different rules, choose the one with the smallest posterior risk.
- That is, take the action according to the rule with the smallest posterior expectation of loss function.
- The resulting rule is called a *formal Bayes rule*.
- Formal Bayes rule: $\delta_0(y)$ is a formal Bayes rule if $r(\delta_0|y) < \infty$ for all y and $r(\delta_0|y) \leq r(\delta|y)$ for all y and δ.
- In theory, this is all we need to know for all sorts of decision problems (e.g., prediction, point estimation, and hypothesis testing).
- For example, as we will see later, if we have a simple 0-1 loss function and a discrete action space such as the coin tossing example, the formal Bayes rule is choosing the mode of the posterior distribution $P(v|y)$.

Squared error loss

- Many decision problems in statistics deal with estimating the parameter of a probability model (e.g., the mean of a normal model, or the coefficients in a linear regression model), i.e. we have $\mathcal{V} = \theta$.

- A possible loss function is the *squared error loss* function: $L(\theta, a) = (\theta - a)^2$.

- In general, the formal Bayes rule for this specific loss function is to choose the mean of the posterior distribution:

$$\begin{aligned} E_{\theta|y}(L(\theta, a)|y) &= E_{\theta|y}((\theta - a)^2|y) = E_{\theta|y}(\theta^2 - 2a\theta + a^2|y) \\ &= E_{\theta|y}(\theta^2|y) - 2aE_{\theta|y}(\theta|y) + a^2 \end{aligned}$$

  We take the the derivative with respect to a and set it to zero:

$$-2E(\theta|y) + 2a = 0 \Rightarrow a = E(\theta|y)$$

- That's the reason we usually use posterior expectation for point estimate.

Absolute error loss

- Now suppose we want to use the *absolute error loss* function: $L(\theta, a) = |\theta - a|$
- Therefore, we need to minimize $E_{\theta|y}(|\theta - a|)$
- Using Leibniz's rule,

$$\frac{\partial}{\partial t}\int_{a(t)}^{b(t)} f(x,t)dx = \int_{a(t)}^{b(t)} \frac{\partial}{\partial t} f(x,t)dx - f(a(t),t)a'(t) + f(b(t),t)b'(t)$$

we have

$$\begin{aligned}\frac{\partial}{\partial a}E_{\theta|y}(|\theta - a|) &= \frac{\partial}{\partial a}\int_{-\infty}^{a}(a-\theta)f(\theta|y)d\theta + \frac{\partial}{\partial a}\int_{a}^{\infty}(\theta - a)f(\theta|y)d\theta \\ &= \int_{-\infty}^{a} f(\theta|y)d\theta - \int_{a}^{\infty} f(\theta|y)d\theta\end{aligned}$$

- This is zero when a is set to the median of the posterior distribution.

Hypothesis testing

- Another type of decision problem, as we mentioned above, is hypothesis testing.
- If we want to choose between two hypothesis $H_0 : \mu \leq 0$ and $H_1 : \mu > 0$, all we need to do again is to choose the hypothesis whose posterior risk is smaller.
- Let's assume a simple $0 - 1$ loss function.
- The posterior risk of accepting null, H_0, (or as it is stated by some statisticians: not rejecting it) is

$$0 \times P(H_0|y) + 1 \times P(H_1|y) = P(H_1|y)$$

- The posterior risk of accepting H_1 is similarly $P(H_0|y)$.
- Therefore, choosing the hypothesis with a smaller posterior risk means choosing the one withe a higher posterior probability.
- This is also the reason we classify objects to the the class with a highest posterior probability in classification models.

Hypothesis testing

- In general the loss due to *type I* error (i.e., rejecting H_0 when it is true) is different from that of *type II* error (i.e., accepting H_0 when it is not true).
- In this case, although we might not choose the one with a higher posterior probability, the principle of choosing the one with a smaller posterior risk remains as before.
- Let's assume the loss due to type I error is 19 and the loss due to type II error is 1.
- We accept H_1 (reject H_0) if its posterior risk is smaller than the posterior risk of H_0, i.e.,

$$\begin{aligned} 0 \times P(H_1|y) + 19 \times P(H_0|y) &< 0 \times P(H_0|y) + 1 \times P(H_1|y) \\ 19P(H_0|y) &< P(H_1|y) \\ P(H_0|y) &< \frac{1}{20} = 0.05 \end{aligned}$$

- That is, for this arbitrary loss function, we reject the null hypothesis if its posterior probability is less than 0.05.

Posterior odds

- Now let's consider a simple hypothesis testing problem formalized as a decision problem between two possible models: $P(y|\theta_0)$ and $P(y|\theta_1)$. That is, we think θ, the parameter of the model, could take one of the two possible values.

- *A priori*, we believe the probabilities of $\theta = \theta_0$ and $\theta = \theta_1$ are $P(\theta_0)$ and $P(\theta_1)$ respectively.

- With a simple 0-1 loss function, we choose the model (i.e., θ) with a higher posterior probability. We could compare posterior probabilities by presenting them in the form of a posterior odds $P(\theta_0|y)/P(\theta_1|y)$ as follows:

$$\frac{P(\theta_0|y)}{P(\theta_1|y)} = \frac{P(\theta_0)P(y|\theta_0)/P(y)}{P(\theta_1)P(y|\theta_1)/P(y)} = \frac{P(\theta_0)P(y|\theta_0)}{P(\theta_1)P(y|\theta_1)}$$

  That is, the posterior odds is the prior odds, $P(\theta_0)/P(\theta_1)$, multiplied by the likelihood ratio, $P(y|\theta_0)/P(y|\theta_1)$.

Bayes factor

- Traditionally, statisticians avoid expressing a prior odds in favor of either alternatives (especially if we are not making a decision, rather, we are reporting our findings): $P(\theta_0)/P(\theta_1) = 1$, and rely only on

$$\frac{P(y|\theta_0)}{P(y|\theta_1)}$$

  which is known as **Bayes factor**.

- This is analogous (not the same in general settings though) to the likelihood ratio test that is commonly used in the frequentist framework.

- When H_0 and H_1 are not single point hypothesis, the Bayes factor is defined in general as

$$BF = \frac{\int p(y|\theta_0)d\theta_0}{\int p(y|\theta_1)d\theta_1}$$

Bayes factor

- We can also use BF to choose between two alternative models:

$$BF_{12} = \frac{P(y|M1)}{P(y|M2)}$$

- In general, when the models are specified in terms of unknown parameters, θ, we have

$$BF_{12} = \frac{\int P(y|\theta_1, M_1)P(\theta_1|M_1)d\theta_1}{\int P(y|\theta_2, M_2)P(\theta_2|M_2)d\theta_2}$$

- That is, BF is the ratio of *prior predictive distributions*.

Bayes factor

- Jeffreys (1961) provided interpretive ranges for the BF analogous to what frequentists use for p-values:
  - $1 < BF < 3$: slight evidence
  - $3 < BF < 10$: positive evidence
  - $BF > 10$: strong evidence
- Using the BF has some difficulties. For example, in general we cannot use improper prior distributions.
- Other alternatives such as fractional Bayes Factor (O'Hagan 1995) are more appropriate (this is beyond the scope of this course, but you can refer to the paper by O'Hagan: "Fractional Bayes factors for model comparison", JRSS, 1995, 56, 99-118).

Model selection based on deviance

- We mentioned that with a 0-1 loss function, we choose the model with a higher posterior probability.
- It turns out (as discussed in Appendix B in Gelman et. al., 2002), in the limit of large sample sizes, the model with the highest posterior probability would have the lowest KL (Kullback-Leibler) information, and as the result the lowest expected deviance.
- Deviance, which is defined as $D(y, \theta) = -2\log(p(y|\theta))$, is a measure of discrepancy (i.e., lack of fit, therefore lower deviance is better).
- For example, for the $y_i \sim \text{binomial}(n_i, \theta)$ model

$$D(y, \theta) = -2\Big\{\sum_i \big[y_i \log\theta + (n_i - y_i)\log(1-\theta) + \log\binom{n_i}{y_i}\big]\Big\}$$

- For the $y_i \sim \text{Poisson}(\theta)$ model

$$D(y, \theta) = -2\Big\{\sum_i \big[y_i \log\theta - \theta - \log(y_i!)\big]\Big\}$$

Model selection based on deviance

- The deviance measure as described above, depends on both y and θ.
- If we want to use a measure that depends only on y, we can integrate the deviance over the posterior distribution

$$D_{avg}(y) = E(D(y, \theta)|y)$$

- We can estimate this by using simulated samples from the posterior distribution

$$\hat{D}_{avg}(y) = \frac{1}{L} \sum_{\ell=1}^{L} D(y, \theta^{\ell})$$

Model selection based on deviance

- Deviance is especially used when we compare nested models; that is, when we are deciding whether to include the predictor x in the model or not, i.e.:

$$\begin{aligned} M_0: &\quad y = \beta_0 + \varepsilon \\ M_1: &\quad y = \beta_0 + \beta_1 x + \varepsilon \end{aligned}$$

- However, we could decrease deviance by arbitrarily increasing the complexity of model, for example, by adding more predictors into the model.

- In general, it is recommended to use more complex models only when they result in substantial (i.e., statistically significant) improvement in performance (i.e, substantial decrease in deviance).

- The above principle is widely known as Occam's razor stating that "entities should not be multiplied beyond necessity", or in simple words: "everything equal, we should use the simplest solution".

Deviance Information Criterion (DIC)

- When we are relying on deviance, we need a measurement that accounts for the trade-off between complexity and goodness-of-fit.
- In a decision model, this could be done by using a loss function that penalizes larger models (i.e., everything equal, we favor simplicity).
- A simple measure, which does this automatically, is called *deviance information criterion* (DIC) defined as follows (Spiegelhalter et. al. 2002):

$$DIC = \hat{D}_{avg}(y) + p_D$$

- p_D is called *effective number of parameters* and is a measure of complexity

$$p_D = \hat{D}_{avg}(y) - D_{\hat{\theta}}(y)$$

Deviance Information Criterion (DIC)

- Here, $D_{\hat{\theta}}(y)$ is the deviance when we first average posterior parameters and then calculate deviance (as opposed to integrating deviance over posterior parameters).

- Therefore, we can obtain DIC as follows:

$$DIC = 2\hat{D}_{avg}(y) - D_{\hat{\theta}}(y)$$

- Caution! Although it is easy to use DIC for model evaluation, remember that the best approach is still to use problem specific loss function, and based on the posterior risk, to find the optimal decision rule. Use DIC only when you don't have a better loss function or you simply want to report your findings.

Example: Titanic survival

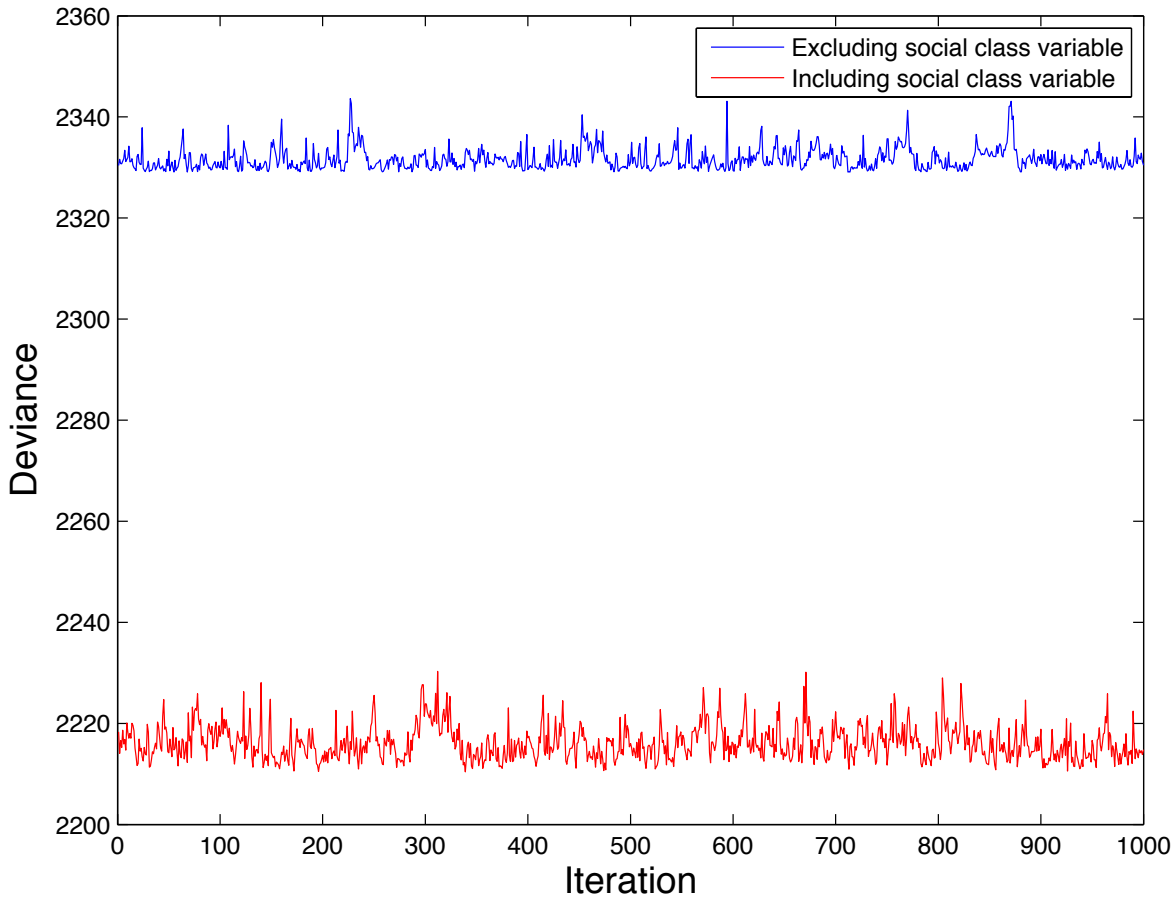
- Recall the Titanic dataset.
- We consider two nested logistic regression models: Model M_0, which does not include the social class predictor (i.e., only the intercept, age and gender are included), and Model M_1, which includes the social class as well as other variables.
- We fit these two models separately and present the results in the following table

Model	$\hat{D}_{avg}$	$D_{\hat{\theta}}$	p_D	DIC
M_0	2331.6	2329.1	2.5	2334.1
M_1	2216.2	2210.1	6.1	2222.4

- As we can see, M_1 has a smaller DIC, and therefore, provides a better fit. This could be interpreted as statistical significance of social class.

Example: Titanic survival

- We can also compare the posterior distribution of D for different models. The following graph shows the trace plot of D for models M_0 and M_1.



Model selection using out-of-sample data

- If our objective for modeling is prediction, it would be more appropriate to decide between alternative models based on their performance on future data.
- While the theory of decision making remains the same as before, in practice it is common to hold out a subset of observed data (i.e., not including it in our modeling) and treat it as a test set in order to estimate posterior risk.
- We should try to use a relevant loss function specific to the problem at hand.
- If we don't have such a loss function, we may use simple loss functions such as 0-1 or squared error.
- In this case, a common measures for evaluating alternative models is the log of posterior predictive probability: $\log(p(\tilde{y}|y))$, where $\tilde{y}$ is the true value of test cases.

Bayesian Asymptotics: Discrete Case

- When the parameter space is finite, $\Theta = \{\theta_0, \theta_1, \ldots, \theta_k\}$ with prior $p_j = P(\theta = \theta_j) > 0$, then given n observed samples from $P(y|\theta_0)$, we can show that

$$\lim_{n\to\infty} P(\theta = \theta_0|y) = 1 \qquad \lim_{n\to\infty} P(\theta = \theta_j|y) = 0 \quad \forall j \neq 0$$

- To see this, consider the log-posterior odds with respect to θ_0 (i.e., true value of model parameter)

$$\log\left(\frac{P(\theta|y)}{P(\theta_0|y)}\right) = \log\left(\frac{P(\theta)}{P(\theta_0)}\right) + \sum_i^n \log\left(\frac{P(y_i|\theta)}{P(y_i|\theta_0)}\right)$$

- For $\theta \neq \theta_0$, the expectation of each summand in the second term is negative so the right hand side approaches $-\infty$ as $n \to \infty$. Therefore,

$$\lim_{n\to\infty} P(\theta = \theta_j|y) = 0 \quad \forall j \neq 0$$

- Because the sum of probabilities is 1,

$$\lim_{n\to\infty} P(\theta = \theta_0|y) = 1$$

Bayesian Asymptotics: Continuous Case

- Doob (1949) showed that in general if a consistent estimator exists for θ, the posterior distribution tends to concentrate near the true value with probability 1 under the joint distribution of the data and parameter.

- We can also show that (given some regularity conditions) for large n, the data overwhelms the prior, and the posterior distribution is asymptotically normal (the proof is similar to what we use for Laplace approximation),

$$\theta|y \overset{\text{asy}}{\sim} N(\hat{\theta}_n, I(\hat{\theta}_n)^{-1})$$

where $\hat{\theta}_n$ is the MLE and $I(\hat{\theta}_n)$ is the observed Fisher information.