10-716: Advanced Machine Learning **Spring 2019**

Lecture 3: January 22

Lecturer: Pradeep Ravikumar *Scribes: Jocelyn Huang, Nicholas Trieu, Shuyang Yang*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

3.1 Decision Rules

Recall that a randomized decision rule takes the form: $\delta^*(x, A) = Pr(A \subseteq \mathcal{A} \text{ is picked})$.

On the other hand, a deterministic rule specifies a single action: $\delta(x) \in \mathcal{A}$.

The question is, can we simplify a randomized decision rule $\delta^*(x, A)$ to a deterministic $\delta(X)$? If we consider $\mathcal{A} \in \mathbb{R}^d$, we can see that we cannot always act on the expected value $\mathbb{E}_{a\sim\delta^*(x,\cdot)}[a]$, as the result may not lie within $\mathcal{A}$ unless $\mathcal{A}$ is convex.

Theorem 3.1 ***Derandomization.*** *Suppose $\mathcal{A} \in \mathbb{R}^d$ is convex, and $L(\theta, a)$ is a convex function of a $\forall\theta \in \Theta$. Let δ^* be a randomized rule such that $\delta = \mathbb{E}_{a\sim\delta^*(x,\cdot)}|a| < \infty$ $\forall x \in \mathcal{X}$. Then, $L(\theta, \delta(X)) \leq L(\theta, \delta^*(x, \cdot))$.*

Proof: We use the definition of our deterministic rule and Jensen's inequality to prove this theorem.

$$L(\theta, \delta(X)) = L(\theta, \mathbb{E}_{a\sim\delta^*(x,\cdot)}[a]) \tag{3.1}$$
$$\leq \mathbb{E}_{a\sim\delta^*(x,\cdot)} L(\theta, a) \text{ by Jensen's inequality} \tag{3.2}$$
$$= L(\theta, \delta^*(X, \cdot)) \tag{3.3}$$

■

Given the conditions of Theorem 3.1, this theorem tells us that there exists a simple rule derandomizer that doesn't sacrifice loss. Note that this does not work for discrete action spaces, only for continuous spaces where the action space is convex.

In the previous lecture, we saw a method of going from a randomized decision rule $\delta^*(X, A)$ to a decision rule that just uses the sufficient statistic t, even if the action space is non-convex: $\delta^*(t, A) = \mathbb{E}_{X|t}\delta^*(X, A)$.

If we try to perform a similar operation on a deterministic decision rule, we get $\delta(t) = \mathbb{E}_{X|t}\delta(X)$, which may not lie in the action space. So as before, this can only be performed when the action space is convex. However, the former, randomized method is well-defined and can also operate on deterministic decision rules, so it is more general and can still be used.

For the specific case where the action space is convex, though, we can also improve a deterministic decision rule using the Rao-Blackwell theorem, as follows.

Theorem 3.2 ***Rao-Blackwell Theorem.*** *Suppose $\mathcal{A} \in \mathbb{R}^d$ is convex, and $L(\theta, a)$ is a convex function of a $\forall\theta \in \Theta$. Let T be a sufficient statistic for θ, and let δ^0 be a deterministic decision rule. Then, the deterministic decision rule $\delta^1(t) = \mathbb{E}_{X|t}[\delta^0(X)]$ is R-equivalent to or R-better than δ^0.*

Proof:

$$\begin{aligned} L(\theta, \delta^1(T)) &= L(\theta, E_{X|T}[\delta^0(X))] \\ &\leq E_{X|T}[L(\theta, \delta^0(X))] \quad \text{pulling the expectation out, and using Jensen's inequality for convex } L(). \\ R(\theta, \delta^1) &= E_T[L(\theta, \delta^1(T))] \\ &\leq E_T[E_{X|T}[L(\theta, \delta^0(X))]] \\ &= R(\theta, \delta^0). \end{aligned}$$

■

The Rao-Blackwell theorem shows that if we have a sufficient statistic and some decision rule, we have an easy way of improving on the decision rule (at the very least, it won't be worse than the original).

3.2 Bayesian Analysis

3.2.1 Definitions

A **posterior distribution** of θ given samples X is defined as:

$$\Pi(\theta|X) = \frac{\Pi(\theta)f(X|\theta)}{\int \Pi(\theta)f(X|\theta)d\theta)} = \frac{\Pi(\theta)f(X|\theta)}{m(X)}$$

Let $\mathcal{F} = \{f(X|\theta)\}_{\theta\in\Theta}$ be a class of density families (i.e. Gaussians). A class $\mathcal{P}$ denoting the prior distribution is **conjugate** to $\mathcal{F}$ if $\Pi(\theta|X) \in \mathcal{P}$.

Example: Let $X \sim N(\theta, \sigma^2), \theta \sim N(\beta, \tau^2)$. Then $\Pi(\theta|X) = N(\mu(x), \rho^2)$ where

$$\begin{aligned} \mu(x) &= \frac{\tau^2}{\tau^2+\sigma^2}X + \frac{\sigma^2}{\tau^2+\sigma^2}\beta \\ \frac{1}{\rho^2} &= \frac{1}{\sigma^2} + \frac{1}{\tau^2}. \end{aligned}$$

If we instead use the **improper prior** $\Pi(\theta) = 1$ (not a valid probability distribution), we can still apply the Bayes rule and get posterior is $\Pi(\theta|X) = N(X, \sigma^2)$.

So even though the prior may not be a valid probability distribution, the posterior may still be a valid distribution.

A $100(1-\alpha)$ **credible set** C for θ is a subset of θ such that

$$1 - \alpha \leq P(C|X) = \int_C \Pi(\theta|X)\ d\theta.$$

The **size** of C is given by $S(C) = \int_C h(\theta)\ d\theta$, where $h : \Theta \to \mathbb{R}_+$. Notice that when $h(\theta) = 1$, this yields the volume of the set C, a simple and straightforward measure of the size of the set. In choosing a credible sets for θ, it is usually desirable to try to minimize its size. To do this, one should include in the sets only those points with the largest posterior density, i.e., the most likely values of θ. (For example, if we can only pick one element, we would like to pick the one with the largest probability mass.) Our objective is then to find $\inf_C S(C)$ such that $1 - \alpha \leq P(C|X)$.

The solution is given by $K(\alpha)$ where

$$C_K = \{\theta : f(\theta|x) > K\}$$
$$K(\alpha) = \sup\{K : P(C_K) \geq 1 - \alpha\}$$

Essentially, we find sets of elements with the highest probability masses, then find the smallest of those sets that has a posterior probability density at least $1-\alpha$, yielding $C_{K(\alpha)}$ as a solution to the above optimization problem.

3.2.2 Bayesian Analysis and Decision Theory

The strength of Bayesian analysis in decision theory is that it combines the loss function and the prior into something actionable. We start with some definitions.

Bayesian Expected Loss:

$$\rho(\Pi(\theta|x), a) = \int_\Theta L(\theta, a)\Pi(\theta|x)\ d\theta.$$

So now, we have some actionable metric that gives us the Bayes estimator in two stages:

1. We compute $\Pi(\theta|x)$ using the Bayes rule, combining prior information with the information that we have gained from the sample.

2. We can take this posterior distribution and pick an action based on the loss function that minimizes the Bayesian expected loss.

This is called the Bayes estimator. More formally:

Bayes Estimator:

$$\delta^\Pi(x) = \inf_{a \in \mathcal{A}} L(\Pi(\theta|x), a).$$

Another quantity that one might be concerned about is the Bayes risk, which takes the expectation of the loss over the possible distributions of θ, i.e. all possible states of nature.

Bayes Risk:

$$r(\Pi, \delta) = E_{\theta \sim \Pi}[E_{x \sim f(x|\theta)}[L(\theta, \delta(x))]].$$

Clearly, we would like to minimize the Bayes risk as well, to get a decision rule $\delta^\Pi \in \inf_\delta r(\Pi, \delta)$. But it turns out that minimizing the Bayes risk and the Bayesian expected loss is the same! They are both the Bayes estimator.

Theorem 3.3 ***Minimization Equivalence Theorem.*** *$\delta^\Pi \in \arg\inf_\delta r(\Pi, \delta)$ minimizes $\delta^\Pi(X) \in \arg\inf_{a \in \mathcal{A}} \rho(\Pi(\theta|X), a)$ for $X; m(x) > 0$.*

Proof: First recall the joint distribution: $f(X|\theta)\Pi(\theta) = \Pi(\theta|X)m(X)$.

Then:

$$r(\Pi, \delta) = \int_\Theta \int_\mathcal{X} l(\theta, \delta(X))\ f(X|\theta)\ \Pi(\theta)\ dX d\theta \tag{3.4}$$

$$= \int_\mathcal{X} \int_\Theta l(\theta, \delta(X))\ \Pi(\theta|X)\ m(X)\ dX \tag{3.5}$$

$$= \int_\mathcal{X} \rho(\Pi(\theta|X), \delta(X))\ m(X)\ dX \tag{3.6}$$

■

So we can see that the minimization problems for the Bayesian expected loss and the Bayes risk are the same, and we get the same estimator. This is quite convenient, as the latter formulation may be harder to solve, as it deals with not only the distribution of X but with the distribution over states of nature as well.

3.2.3 Advantages and Criticisms of Bayesian Analysis

Advantages:

1. Incorporates prior information over Θ.
2. $\Pi(\theta|x)$ quantifies the uncertainty over θ via a probability distribution. The expected value is a good way to quantify uncertainty.
3. Conditional perspective.
4. Incorporates loss functions in estimator.

Criticisms:

1. Priors need not be objective, different priors can lead to different answers.
2. Computational difficulty with large-scale priors.

3.3 Minimax Analysis Introduction

Minimax analysis is a frequentist method uses no prior information on Θ. It is useful in the case you want to be conservative with respect to the loss you might suffer, or if nature is adversarial in how it sets parameters. However, it is not very actionable in general.

Minimax analysis stems from game theory, and as such we model actions as part of a two-player game between nature (choosing $\theta \in \Theta$) and ML/a statistician (choosing $a \in \mathcal{A}$). As always, the statistician's goal is to minimize the loss $L(\theta, a)$, while nature's goal is to maximize the loss.

Depending which "player" goes first, i.e. whether the parameter or the decision rule is set first, we define two values:

$$\overline{V} \equiv \inf_{a \in \mathcal{A}} \sup_{\theta \in \Theta} L(\theta, \delta) \quad \text{if nature goes first} \tag{3.7}$$

$$\underline{V} \equiv \sup_{\theta \in \Theta} \inf_{a \in \mathcal{A}} L(\theta, \delta) \quad \text{if the statistician goes first} \tag{3.8}$$

Intuitively, it makes sense that whoever goes first has the advantage, such that $\overline{V} \geq \underline{V}$. We can also show this more formally.

Lemma 3.4 *For $\overline{V}$ and $\underline{V}$ as defined above, $\overline{V} \geq \underline{V}$.*

Proof: By definition of infimum:

$$\sup_{\theta} \inf_{a} L(\theta, a) \leq \sup_{\theta} L(\theta, b) \; \forall b$$

Then, taking an infimum over b on both sides:

$$\inf_{b} \sup_{\theta} \inf_{a} L(\theta, a) \tag{3.9}$$

$$= \sup_{\theta} \inf_{a} L(\theta, a) \leq \inf_{b} \sup_{\theta} L(\theta, b) \tag{3.10}$$

Which we can then see is just $\underline{V} \leq \overline{V}$. ■

The **minimax strategy** for the statistician is $\delta^M \in \arg\inf_{\delta} \sup_{\theta} L(\theta, \delta)$.

The **maximin strategy** for nature is $\theta^M \in \arg\sup_{\theta} \inf_{a} L(\theta, a)$.