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

Lecture 2: January 17

Lecturer: Pradeep Ravikumar *Scribes: Austin Dill, Hima Tammineedi, Conor Igoe*

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

2.1 Review

In the last lecture we started off looking at decision theory, and with that the Bayesian perspective and the Frequentist perspective. We introduced *Bayes Expected Loss*, giving us a way to quantify how good an action is under some definition of loss and given some distribution on the state of nature at decision time. We then discussed *Frequentist Risk* which describes the expected loss using a particular decision rule and supposing a particular state of nature. We also introduced *Bayes Risk*, in a sense combining the two by taking an expectation of *Frequentist Risk* over a distribution on the state of nature.

So far however, these perspectives alone are not sufficient for a particular decision problem. How should we pick a decision rule? We briefly discussed *Admissibility* in the previous lecture to restrict our search, and in this lecture we go further, introducing the major principles for actually making a decision or choosing a rule.

2.1.1 Decision Theory Setup

Recall:

Θ : Parameter Space
$\mathcal{X}$: Sample Space
$\mathcal{A}$: Set of Actions
$L : \Theta \times \mathcal{A} \to \mathbf{R}$ Loss function
$\delta : \mathcal{X} \to \mathcal{A}$ Decision rule (deterministic)
π : distribution on $\theta \in \Theta$

(Conditional) Bayesian Expected Loss	**(Frequentist) Risk**	**Bayes Risk**
$\rho(\pi, a) = \mathbb{E}_\pi L(\theta, a)$	$\mathrm{R}(\theta, \delta) = \mathbb{E}_X L(\theta, \delta(X))$	$r(\pi, \delta) = \mathbb{E}_\pi R(\theta, \delta)$

Note that Risk is a Frequentist notion as we are explicitly taking an expectation over all possible samples, holding θ fixed. However, this is merely a perspective – it is not a principle in the sense that it does not really tell you how to come up with a δ. Some Frequentist principles that we'll discuss today are more actionable.

On the other hand there is the Bayesian perspective which essentially says that, instead of fixing θ and averaging over X, you fix X and average over θ. We also had a corresponding principle in the Bayesian

perspective, which was to use the conditional Bayesian expected Loss, with actions chosen so as to minimize this quantity.

2.1.2 Bayes Risk Example

$$\begin{aligned} \theta &\in \mathbf{R} \equiv \Theta \\ X &\sim N(\theta, \sigma^2) \\ \theta &\sim N(0, \tau^2) \qquad \text{(prior } \pi\text{)} \\ L(\theta, a) &= (\theta - a)^2 \end{aligned}$$

Consider the following class of decision rules parameterized by $c \in \mathbf{R}$:

$$\begin{aligned} \delta_c(X) &= cX \\ \text{Bayes Risk: } r(\pi, \delta_c) &= c^2 + (1-c)^2\tau^2 \end{aligned}$$

Using the Bayes Risk we can obtain the Bayes Decision Rule by minimizing:

$$\begin{aligned} 2c + 2(c-1)\tau^2 &= 0 \\ \implies c &= \frac{\tau^2}{1+\tau^2} \end{aligned}$$

Observe that our prior is influencing our estimate of θ by pulling our estimate closer towards zero based on how "confident" our prior belief is (variance τ^2).

2.2 Minimax Principle

We will now consider a worst-case analysis approach to decision theory. Under this scheme, we will prefer δ_1 to δ_2 if

$$\sup_{\theta \in \Theta} R(\theta, \delta_1) < \sup_{\theta \in \Theta} R(\theta, \delta_2)$$

Minimax Rule:

$$\delta_m^* \in \arg\inf_{\delta} \sup_{\theta \in \Theta} R(\theta, \delta)$$

Example: Suppose $X \sim N(\theta, 1)$ and $\delta_c = cx$.

$$R(\theta, \delta_c) = c^2 + (1-c)^2\theta^2$$

$$\sup_{\theta\in\Theta}\left\{c^2+(1-c)^2\theta^2\right\}=\begin{cases}1 & c=1\\ \infty & c\neq 1\end{cases}$$

Therefore, δ_1 is minimax.

2.3 Conditional and Frequentist Approaches

In the next couple examples, we will explore the difference in outcomes between the conditional and frequentist perspective.

Example: Let $P(X=\theta-1)=P(X=\theta+1)=\frac{1}{2}$ and that $X_1, X_2 \sim P$. Let $\mathbf{X}=(X_1,X_2)$.

Let us consider the loss function L(θ,a) $=\mathbb{I}(\theta\neq a)$

A reasonable decision rule might be the following:

$$\delta(\mathbf{X})=\begin{cases}(X_1+X_2)/2, & X_1\neq X_2\\ X_1-1, & X_1=X_2\end{cases}$$

Conditional Perspective: Since we know whether we are in the case where X_1 and X_2 are equal or not, we know if we are in a situation where our probability of error will be 0% or 50% respectively.

Frequentist Perspective: In a frequentist model, we must average over all possible data we could have drawn from the distribution. This will give us a risk of .25 under a 0-1 loss.

Example: Let $\mathcal{X}=\{1,2,3\}$ and $\Theta=\{0,1\}$ with X distributed as follows:

	x		
	1	2	3
$f(x\|0)$	0.005	0.005	0.99
$f(x\|1)$	0.0051	0.9849	0.01

Frequentist Perspective: A reasonable decision rule might be the following,

$$\delta(X)=\begin{cases}0, & X=3\\ 1, & X=1\vee 2\end{cases}$$

Under a 0-1 loss, $R(\theta,\delta)=P(\delta(X)\neq\theta)$. Therefore, $R(0,\delta)=0.01$ and $R(1,\delta)=0.01$.

Conditional Perspective: Say $x=1$. The error should be closer to a fifty-fifty chance because the likelihoods are so close. This implies that Frequentist Risk can be particularly misleading.

2.4 Likelihood Principle

So far we have seen many principles that are concerned with *possible* data samples that could be seen. The Likelihood Principle makes central the idea that only the actually observed data x should be relevant to considerations about θ.

Definition 2.1 *Let $l(\theta) = f(x|\theta)$ and consider this a function of θ for fixed x. As this is based on the density of x, it will be called the likelihood function.*

Note that this can still be viewed from a Frequentist lens because we consider the true underlying θ to be fixed - ie. we condition on θ, instead of averaging over it.

The Likelihood Principle: When making decisions about θ after observing x, all relevant experimental information about θ is contained in the likelihood function and if two likelihood functions are proportional, they contain the same information about the state of nature.

Next, we'll consider some examples that shows the power of The Likelihood Principle and the difference in results achieved when using it.

Example: Let E_1 and E_2 be two experiments conducted on the same spaces $\mathcal{X} = \{1, 2, 3\}$ and $\Theta = \{0, 1\}$. These experiments share a common θ and consist of observing X_1 and X_2 respectively with the following densities:

	x_1		
	1	2	3
$f_1(x_1\|0)$	0.90	0.05	0.05
$f_1(x_1\|1)$	0.09	0.055	0.855

	x_2		
	1	2	3
$f_2(x_2\|0)$	0.26	0.73	0.01
$f_2(x_2\|1)$	0.026	0.803	0.171

Even though these experiments look very different in terms of their probabilities, under the likelihood principle they provide identical information. Consider the case where $x_1 = 1$ or $x_2 = 1$ is observed. As the ratio between $f_1(1|0)$ and $f_1(1|1)$ is the same as the ratio between $f_2(1|0)$ and $f_2(1|1)$ the evidence for θ is the same in these two experiments.

From a classically frequentist perspective, you can create decision rules that produce drastically different results for the above experiments.

Example: Imagine a classical frequentist statistician is approach by an engineer who has been collecting data with his voltmeter and would like to draw statistical conclusions from his experimentation.

The engineer has collection 100 observations ranging from 75 to 90 volts with a mean of 87 and a standard deviation of 4. The statistician may suggest a model that each observation is IID from a $N(\theta, \sigma^2)$ distribution.

Later, the engineer returns and says that the voltmeter actually was broken and that it truncated any measurement over 100 volts. Intuition would imply that this should affect the conclusions drawn from the data, but the statistician concludes that a new model must be devised and different conclusions drawn. They therefore settle on a model where $Z \sim N(\theta, \sigma^2)$ and $X = \max(100, Z)$.

Again, the engineer returns and says there was actually has a second voltmeter that was working correctly and would have used it if any measurement at exactly 100 was recorded. This allows the statistician to revert to the original model.

If the engineer then revealed that the second voltmeter was also inoperable they would have to revert to the more complicated model.

Takeaway: The Likelihood Principle states that the difference between these experimental setups shouldn't matter because we live in a world where the data collected had no points at or above 100 volts. This corresponds to our natural intuition.

(Weak) Conditional Principle: When there are multiple experiments that could have been performed, only the experiment that was actually perform should contribute information about the state of nature, θ/

2.5 Sufficiency Principle

• The (weak) conditional principle + the sufficiency principle ⇒ likelihood principle

---

• Sufficient statistic: A function of the data that summarizes all the information you need to infer θ given the model.

Example:

$$x_i \sim \mathcal{N}(\mu, \sigma^2),\ i = 1, \ldots, n$$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

We claim that $\overline{x}, \overline{\sigma^2}$ are sufficient statistics of the data for you to infer μ, σ^2.
We have condensed n numbers to just two, but this is actually still sufficient information to re-derive the original parameters.

Let X be a RV whose distribution depends on θ.
We say T (a function of $X \in \mathcal{X}$) is a sufficient statistic of θ IFF $X|T(X) = t$ is independent of θ.

We also define the set of all sufficient statistics (i.e. the range of T): $\mathcal{T} = \{T(X) : X \in \mathcal{X}\}$.

Definition 2.2 *The partition induced by T is $X_t = \{X : T(X) = t\}$.*

$X|T(X) = t$ has domain X_t and the density $f_t(X)$ does not depend on θ.
Also note that for t $\neq t'$: $\bigcup \mathrm{X_t} = \mathrm{X}$ and $\mathrm{X_t} \bigcap \mathrm{X_{t'}} = \phi$

• We define a 2 step process to generate X:

1. Generate t with some distribution depending on θ.
2. Pick $X \sim f_t(\cdot)$ independent on θ. (ie. the density within each component does not depend on θ.)

Theorem 2.3 *Factorization Theorem: $f(X|\theta) = g(T(X), \theta) \cdot h(X)$ IFF T is a sufficient statistic of θ*

We can see that the right-hand side of the equation in the factorization theorem matches the generator process we defined above.

Theorem 2.4 *Let $\delta_0^*(X, \cdot)$ be a randomized decision rule and let T be a sufficient statistic of θ.*
Then, $\exists$ randomized decision rule $\delta_1^*(t, \cdot)$ that only depends on $T(X) = t$ (the sufficient statistic) which is R-Equiv to $\delta_0^*(X, \cdot)$.

Proof: Consider $A \in \mathcal{A}$, $t \in \mathcal{T}$
$\delta_1^*(t, \cdot) = \mathbb{E}_{X|T(X)=t}\delta_0^*(X, A)$
This is an expectation within a particular component of the partition.
Note that the $\delta_0^*(X, A)$ here is the probability of choosing an action in A based on X.

Question: Why don't we use $\delta_1^*(t) = \mathbb{E}_{X|T(X)=t}\delta_0^*(X)$?
Answer: We do not because this expectation need not lie in $\mathcal{A}$ since we aren't given that set is convex. So

instead, we take an average of the probabilities of the actions instead of an average of the actions themselves in order to ensure that the result is well-defined.

—

$$
\begin{aligned}
L(\theta, \delta_1^*(t,\cdot)) &= \mathbb{E}_{a\sim\delta_1^*(t,\cdot)} L(\theta, a) \\
&= \mathbb{E}_{X|t}\ \mathbb{E}_{a\sim\delta_0^*(X,\cdot)} L(\theta, a) \\
\textit{Thus,}& \\
R(\theta, \delta_1^*) &= \mathbb{E}_T\ L(\theta, \delta_1^*(t,\cdot)) \\
&= \mathbb{E}_T\ \mathbb{E}_{X|T}\ \mathbb{E}_{a\sim\delta_0^*(X,\cdot)} L(\theta, a) \\
&= \mathbb{E}_X\ \mathbb{E}_{a\sim\delta_0^*(X,\cdot)} L(\theta, a) \qquad \text{(by the law of total expectation on the outer two expectations)} \\
&= R(\theta, \delta_0^*)
\end{aligned}
$$

Thus, we have shown that δ_1^* is R-Equiv to δ_0^*. ■

An interesting fact is that even when δ_0^* is not randomized, δ_1^* is still randomized.
Proof: $\delta_1^*(t, A) = \mathbb{E}_{X|t}\ \mathbb{I}(\delta_0^*(X) \in A) = P_{X|t}(\delta_0^*(X) \in A)$