Statistical Decision Theory
10716: Advanced Machine Learning
Pradeep Ravikumar

1 Preliminaries

The field of statistical machine learning broadly seeks to answer the question: how can we come up with an inductive system, or a "learning procedure" that automatically improves with experience? While this seems like a very broad question, we can formalize this mathematically using the language of statistical decision theory. Given a suitable statistical notion of task, a loss function to measure performance in the task, and a suitable statistical notion of experience, one can then ask for "optimal" procedures, wrt the loss function, that incorporate finite experience. As we can see, this has a number of degrees of freedom, and understanding this from first principles is the goal of this lecture; and which forms the sub-field of statistical decision theory [Berger, 2013].

Let $\theta \in \Theta$ denote the "state of nature". This state of nature is tied to our formalization of the "task" we are considering, but for now let us think about this abstractly. We do not observe this state of nature, instead, we observe a random variable $X \in \mathcal{X}$ with distribution $P(\cdot; \theta)$ specified by the state of the nature. This random variable represents the data or "experience" that we observe. Our goal is to figure out some aspect of the state of nature (or perhaps simply the state of nature itself) just given the random sample X. Let $\mathcal{A}$ denote the set of possible outputs, typically this will simply be Θ when we want to estimate the entire state of nature.

Let $\delta : \mathcal{X} \mapsto \mathcal{A}$ be an estimator. How good is this estimator? To answer this, we need the notion of a loss function $L : \Theta \times \mathcal{A} \mapsto \mathbb{R}$, so that $L(\theta^*, a)$ quantifies the cost of estimate $a \in \mathcal{A}$ when the state of nature is θ^*. While this quantifies the cost of an *estimate* $a \in \mathcal{A}$, we can also use this to quantify the cost of an *estimator* δ as:

$$R(\theta^*, \delta) = \mathbb{E}_{X \sim P(\cdot; \theta^*)} L(\theta^*, \delta(X)).$$

This quantity is also called the **risk** of the estimator δ. Now that we can evaluate the goodness of any estimator, we can then ask: what is the best possible estimator? This is hard to answer in general, because we may have two estimators δ_1, δ_2 neither of which is dominated with respect to risk by the other, so that there exist states of nature θ_1, θ_2 such that $R(\theta_1, \delta_1) < R(\theta_1, \delta_2)$ but $R(\theta_2, \delta_1) > R(\theta_2, \delta_2)$. Since we don't know whether θ_1 or θ_2 is the true state of nature, which estimator do we pick?

Once we fix a state of nature θ^*, then there is a very simple estimator $\delta(\cdot)$ that is optimal, namely: $\delta(X) = \arg\inf_a L(\theta^*, a)$, which typically incurs zero risk at that state of nature θ^*.

It of course does very poorly for other states of nature θ' that are very different from θ^*, but it is optimal for θ^*.

To circumvent this, we would need to define more global notions of optimality. We shall focus on three notions: minimax optimality, and Bayesian optimality, and uniform optimality.

Example. Suppose $\theta^* \in \mathbb{R}^p$, and $X \sim N(\theta^*, \sigma^2 I)$. Consider the loss function $L(\theta, \theta') = \|\theta - \theta'\|_2^2$. Then, given any estimator δ, we can compute the risk:

$$R(\theta^*, \delta) = \mathbb{E}[\|\delta(X) - \theta^*\|_2^2],$$

which is simply the mean squared error.

2 Minimax Risk and Estimators

One global notion of optimality is to take the conservative or worse-case route, which is called minimax optimality. Let Γ be a set of candidate estimators (perhaps the set of all possible (measurable) estimators). Then, the minimax risk wrt Γ is specified as:

$$r(\Theta, \Gamma) := \inf_{\delta \in \Gamma} \sup_{\theta^* \in \Theta} R(\theta^*, \delta),$$

and any estimator δ_{MM} that achieves this minimax risk is said to be a minimax optimal estimator. The main caveats with this notion are practical. For one, it is typically difficult to certify that a given estimator is minimax — indeed, this forms the subject of many 50 page papers. Moreover is also not practically constructive: it requires solving a min-max problem over all candidate estimators, which is typically intractable. There are also criticisms of this notion as being overly conservative: if δ_1 is much much better than δ_2 for most states of nature θ, except for one θ' where it is marginally worse, minimax-optimality might well pick δ_2.

The minimax risk is associated with the so-called minimax principle:

Minimax Principle: An estimator δ_1 is preferred to another estimator δ_2 if its worst case risk is lower: $\max_{\theta^* \in \Theta} R(\theta^*, \delta_1) < \max_{\theta^* \in \Theta} R(\theta^*, \delta_2)$.

The minimax optimal estimator is that which achieves the minimum worst case risk, though this term is used even for estimators that achieve the minimax risk upto some absolute constants that do not depend on key quantities such as the sample size, or problem dimension or complexity parameters.

3 Bayesian Risk and Estimators

Given a prior π over states of nature, we can define the Bayesian risk:

$$r(\pi, \delta) = \int_{\theta^* \in \Theta} R(\theta^*, \delta)\, \pi(\theta^*) d\theta^*,$$

and the estimator δ_π minimizing this Bayesian risk,

$$\delta_\pi \in \arg\inf_\delta r(\pi, \delta),$$

is said to be the optimal Bayesian estimator given prior π.

The Bayesian risk is associated with the so-called Bayes risk principle:

Bayes Risk Principle: An estimator δ_1 is preferred to another estimator δ_2 if its Bayes risk is lower: $r(\pi, \delta_1) < r(\pi, \delta_2)$.

The Bayes estimator is that which achieves the Bayes risk. A related principle is the so-called **conditional Bayes principle**, which given samples X entails choosing a decision or action $a \in \mathcal{A}$ which minimizes

$$\rho(P(\theta^*|X), a) := \int_{\theta^* \in \Theta} L(\theta^*, a) P(\theta^*|X) d\theta^*,$$

where $P(\theta^*|X)$ is the posterior distribution of the state of nature θ^* given the samples X.

It can be seen that the Bayes risk principle and the conditional Bayes principle yield the same answers since $r(\pi, \delta) = \int_X \rho(P(\theta^*|X), \delta(X)) P(X) dX$, which can be minimized over estimators $\delta(\cdot)$ by, for each sample X, setting $\delta(X)$ to the minimizer of $\rho(P(\theta^*|X), a)$ which is precisely the conditional Bayes Principle.

Unlike the minimax case, the Bayes risk is easier to evaluate, or at least approximate, and moreover it is also easier to compute or at least approximate the optimal Bayesian estimator δ_π, by solving for the conditional Bayesian risk:

$$\delta_\pi(x) \in \arg\inf_{a \in \mathcal{A}} \int_{\theta^* \in \Theta} L(\theta^*, a)\, \pi(\theta^*|x) d\theta^*,$$

which unlike the minimax optimal case is a more tractable, typically even a finite dimensional estimation problem. The main caveat is that it requires the specification of a prior π, which essentially specifies the linear combination weights of how to combine the risks at different states of nature; and moreover such a linear combination might not capture the true notion of global risk. Notwithstanding the concerns with the specification of the prior, for medium to higher dimensional problems, the computations above again get intractable, so that this is not always a practical estimator for many modern data settings.

4 Conditionality and Likelihood Principles

So far we have discussed the expected risk $R(\theta^*, \delta)$ as a very natural object of study. It will now be instructive to consider the criticism of thos notion of risk itself: that it does not evaluate an estimator in light of the given set of observations; rather it computes its expected performance over all possible sample sets. This global performance might not be indicative of the local performance given the specific set of observations. This is best illustrated by the following example. Suppose the state of nature is $\theta \in \mathbb{R}$, given which the observation $X \in \mathbb{R}$ has the following distribution: $P_\theta(X = \theta + 1) = P_\theta(X = \theta - 1) = 1/2$. Suppose we are interested in estimating the state of nature θ^*, so that the action space $\mathcal{A} = \Theta$, and that we have the zero-one loss so that $L(\theta^*, \theta) = \mathbb{I}[\theta^* \neq \theta]$. Suppose we see two samples $X = (X_1, X_2)$ from $P(\cdot; \theta)$, and that the estimator is given by: $\delta(X) = \frac{1}{2}(X_1 + X_2)\mathbb{I}(X_1 \neq X_2) + (X_1 - 1)\mathbb{I}(X_1 == X_2)$.

Its risk is then given by $R(\theta^*, \delta) = \mathbb{P}[\delta(X) \neq \theta^*] = 0.25$, for all $\theta^* \in \Theta$. Let $E(X) = \mathbb{I}[X_1 \neq X_2]$ be the event that the two samples are distinct. It can then be seen that conditioned on $E = 1$, the risk of the estimator is zero, since it necessarily is then the case that $\theta^* = (X_1 + X_2)/2$. While conditioned on $E = 0$, the risk of the estimator is 0.5, since θ^* could be either of $X_1 - 1$ or $X_1 + 1$ with equal chance. Thus, a global or unconditional risk $R(\theta^*, \delta)$ of 0.25 is is misleading in both these cases, especially so when the observed sample X is such that $E(X) = 1$, and the estimator is actually always correct. This leads to the so-called conditionality principle, a weaker and easier stated version of which is as follows:

(Weak) Conditionality Principle. In order to estimate the state of nature θ, suppose we can perform two experiments E_1 or E_2. Suppose J is a binary random variable, such that $J = 1$ or 2 with equal probability of $1/2$. Consider the mixed experiment, where we first sample the value of J, and then perform the experiment E_J. Then the information about θ obtained from the mixed experiment E_J should only depend on the experiment E_j that is actually performed.

This is also illustrated by the following example. Suppose that an engineer uses a voltmeter that makes 5 observations each ranging from 75 to 99 volts. He then asks the statistician to estimate the true voltage. The next day, the engineer says, by the way, it seems the voltmeter truncates voltages at 100, but that should not matter since all measured voltages were below 100. The statistician however is worried: not so fast, he says. This changes the distribution of the observations (since it is a truncated random variable), and hence the risk measure computations, and therefore he will have to redo his calculations. Wait, says the engineer, I should let you know I always carry a backup bulkier voltmeter that I use when I see observations equal to 100 (which might indicate truncation), but since I didn't see any values that were equal to 100, I didn't use it. Phew, says the statistician, I don't have to redo my calculations after all. The next day the engineer says, sorry again, it seems the backup bulkier voltmeter is not working, likely for the past many months, but if I had had

to use it and find out that it was broken I would have fixed it. Oh, says the Statistician with a sigh, we now have a truncated distribution again, so I have to redo the calculations after all.

In this example, the statistician with his focus on computing the expected risk is violating the conditionality principle, but also our common sense or rationality intuitions that the estimator should not have to worry about the observations being truncated at 100 conditioned on the fact that the observations were all lower than 100.

But how to operationalize this conditionality principle? The biggest advance towards this was the likelihood principle. Recall the definition of the likelihood function $\ell(\theta) = P(X|\theta)$, which is the density of the observed samples given state of nature θ, as a function of $\theta \in \Theta$.

Likelihood Principle. In making inferences about the state of nature θ after observing X, all relevant information is contained in the likelihood function $\ell(\theta) = P(X|\theta)$.

The likelihood principle was advocated in the 1950s by R. A. Fisher, and G. A. Barnard. Its importance has been bolstered by technical arguments such as Birnbaum [Berger and Wolpert, 1988], who show that it is implied by the weak conditionality principle, and the sufficiency principle (which requires that estimators should be functions of sufficient statistics of the state of nature or parameters thereof). It is to be noted that the likelihood principle by itself is not actionable: in the sense that it is not clear how to construct an estimator that satisfies the likelihood principle. The Bayes estimator is one estimator that does follow the likelihood principle. When we wish to estimate the state θ itself, then the MLE also satisfies the likelihood principle, as we will see presently. But for more general parameters, is is not clear how to satisfy the likelihood principle, and in particular, how to reconcile the likelihood function with the loss function $L(\theta^*, a)$.

5 Uniform Optimality/PAC Principle

While the minimax and Bayesian optimality notions seem the most natural global optimality notions, and indeed occupy most of the mind-space of statistical decision theorists, as noted above, these are typically not practical. Moreover, these are not the typical class of estimators used in practice in statistical ML. What are the classes of estimators typically used in statistical ML, and are they principled from a statistical decision theory standpoint?

Towards this, let us first define a simple if seemingly impossible notion of **uniform optimality**. We say that a decision rule $\widehat{\delta}$ is uniformly optimal if $\forall \theta^* \in \Theta, \forall a \in \mathcal{A}$,

$$L(\theta^*, \widehat{\delta}) \leq L(\theta^*, a).$$

The reason this impossible as is because we know that the constant rule: $\delta(X) = \arg\inf_a L(\theta^*, a)$

would incur minimum typically zero loss. So the above entails, that a uniformly optimal rule should have loss $L(\theta^*, \widehat{\delta}) = 0$, for all θ^*, which is obviously too strong. This is even more so since $\delta(X)$ is random, and hence at most we could expect this to hold with some probability. But also due to information theoretic reasons, we also expect to incur a small amount of non-zero error, so that we would expect to have probably approximately uniformly optimal (correct) estimators. We say that an estimator δ is $\epsilon - \gamma$ probably approximately uniformly optimal if with probability at least $1 - \gamma$, $\forall \theta^* \in \Theta$,

$$L(\theta^*, \widehat{\delta}) \leq \inf_a L(\theta^*, a) + \epsilon.$$

This is simply the PAC (Probably Approximately Correct) formalism, for general decision theoretic settings. These are also called distribution-free bounds, because ϵ, γ do not depend on the specific θ^*. More generally, we might expect distribution-specific bounds: with probability $1 - \gamma_{\theta^*}$, it holds that:

$$L(\theta^*, \widehat{\delta}) \leq \inf_a L(\theta^*, a) + \epsilon_{\theta^*},$$

We will revisit these bounds later on, but for now, how can we expect to get such bounds for arbitrary loss functions L? Is there an estimation principle that is guaranteed to do so? We could use the "generative model" approach, where we first fit a statistical model to estimate the state of nature $\widehat{\theta}$, for instance via the MLE, and then estimate $\arg\inf_a L(\widehat{\theta}, a)$. When using the MLE, this does yield tight bounds for reasonable loss functions, but this has a couple of caveats. The bounds are tight only for reasonable loss functions, and for small enough statistical models. It also requires that we know the statistical model $P(X|\theta)$. As we will see it is be possible to completely eschew the generative model approach, and come up with "distribution-free" decision rules that make *no assumptions* on the generative model $P(X|\theta)$. This is particularly useful for complex modern data where we do not necessarily want to make stringent assumptions on the statistical model. Another crucial caveat is that it might be computationally more expensive to first estimate the entire state of nature, rather than directly predict the optimal action.

As we will see in the next section, there is a very simple estimator, frequently used in modern statistical machine learning that does not have these drawbacks of the MLE.

6 Decomposable Losses and ERMs

So the critical question is how to construct an estimator that is near uniformly optimal with respect to a given decision theoretic loss? One approach is to approximate the loss function $L(\theta^*, a)$ just using samples from P_{θ^*}, to get a surrogate $\widehat{L}(\theta^*, a)$ and then find the optimal action with respect to this. But how to compute such a surrogate function, given that we do not know θ^*, and without explicitly fitting a generative model (e.g. via the MLE) to fit θ^*?

As it turns out, there is a specific sub-class of decision-theoretic loss functions, that we will call **decomposable loss functions**, which are indeed simple to estimate given samples.

Definition 1 *A loss function $L : \Theta \times \mathcal{A} \mapsto \mathbb{R}$ is said to be decomposable iff:*

$$L(\theta^*, a) = \mathbb{E}_{X \sim P(\cdot;\theta^*)} \ell(X, a), \forall \theta^* \in \Theta, a \in \mathcal{A},$$

for some loss function $\ell : \mathcal{X} \times \mathcal{A} \mapsto \mathbb{R}$.

Thus, for decomposable decision-theoretic loss functions, we can express the loss $L(\theta^*, a)$ as the expectation of a term $\ell(X, a)$ that is entirely as a function of the sample X rather than the unknown state of nature θ^*. A key advantage to this is that we could then compute the so-called empirical loss:

$$\widehat{L}_n(\theta^*, a) = \frac{1}{n} \sum_{i=1}^{n} \ell(X^{(i)}, a),$$

entirely using samples $\{X^{(i)}\}_{i=1}^n \sim \mathcal{P}(\cdot; \theta^*)$.

In what might perhaps be confusing terminology from a statistical decision theory standpoint, this empirical loss is typically called empirical risk in ML. Which then leads to the class of **empirical risk minimizers**:

$$\begin{aligned} \widehat{\delta}_{\text{erm}} &:= \arg\min_{a \in \mathcal{A}} \widehat{L}_n(\theta^*, a) \\ &= \arg\min_{a \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} \ell(X^{(i)}, a). \end{aligned}$$

We term such "empirical risk minimizers" that point-wise minimize an empirical surrogate of the loss as satisfying the **empirical loss principle**. This is to distinguish from other estimation principles such as the likelihood, minimax, or Bayesian estimation principles.

The caveat of course is that they are only applicable to decomposable decision-theoretic loss functions. We next look at some examples of decomposable loss functions, and empirical risk minimizers.

6.1 Examples

MLE. Consider a family of distributions $\{P_\theta\}_{\theta \in \Theta}$. Suppose we are given samples $\{X_i\}_{i=1}^n \sim P_{\theta^*}$ for some $\theta^* \in \Theta$, and we wish to estimate θ^* given the n samples. Here the decision or action space $\mathcal{A} = \Theta$. Suppose $L(\theta^*, \theta) = KL(P_{\theta^*}, P_\theta)$. This can be seen to be decomposable since: $KL(P_{\theta^*}, P_\theta) = \mathbb{E}_{X \sim P_{\theta^*}} \log P_{\theta^*}(X)/P_\theta(X)$.

The corresponding empirical risk minimizer with respect to this loss is then given by:

$$\begin{aligned}\widehat{\theta}_n &= \arg\inf_{\theta} \frac{1}{n}\sum_{i=1}^{n} \log P_{\theta^*}(X_i)/P_{\theta}(X_i) \\ &= \arg\inf_{\theta} \frac{1}{n}\sum_{i=1}^{n} -\log P_{\theta}(X_i),\end{aligned}$$

which is the Maximum Likelihood Estimator or MLE. This estimator thus satisfies the empirical loss principle as well as clearly, the likelihood principle. This example also makes clear that the likelihood principle, which entails that estimators only use the likelihood function as a summary of the data, is a special case of the empirical loss principle.

Binary Classification. Let $X \in \mathcal{X}$ denote the so-called input random variable, and $Y \in \{-1, +1\}$ a binary output random variable, jointly distributed as $(X, Y) \sim P \in \mathcal{P}$. Given observations $\{(X_i, Y_i)\}_{i=1}^n \sim P$, we wish to obtain a classifier $f : \mathcal{X} \mapsto \{-1, +1\}$ that minimizes the so-called mis-classification error: $L(P, f) = \mathbb{P}[f(X) \neq Y]$.
Let $\ell(f, (X, Y)) = \mathbb{I}[f(X) \neq Y]$. It can then be seen that $L(P, f) = \mathbb{E}_{(X,Y)\sim P}\ell(f, (X, Y))$, so that the mis-classification error is a decomposable decision-theoretic loss function. The corresponding empirical risk minimizer would then be given as:

$$\begin{aligned}\widehat{f}_n &= \arg\inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell(f, (X_i, Y_i)) \\ &= \arg\inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[f(X_i) \neq Y_i].\end{aligned}$$

The classifier

$$\begin{aligned}f^* &= \arg\inf_{f} L(P, f) \\ &= \text{sign}(P(Y = 1|\cdot) - 1/2)\end{aligned}$$

that minimizes the loss $L(P, f)$ directly is called the Bayes optimal classifier. This might again seem like a terminological mis-step in ML, since this does not seem Bayesian at all: f^* is just minimizing the pointwise loss $L(P, f)$, where the state of nature is $P \in \mathcal{P}$, and the decision or action is $f \in F$.

What might a fully Bayesian treatment look like? Suppose we have some prior π over distributions in $\mathcal{P}$, and given the samples $(X, Y) = ((X_i, Y_i))_{i=1}^n$, we could then compute the posterior $\pi(\cdot|(X, Y))$, and compute the Bayes optimal estimate as:

$$f_{\text{Bayes}} = \arg\min_{f} \rho(\pi(\cdot|(X, Y)), f),$$

which can be seen to be much more complicated than what is called the Bayes optimal classifier in ML.

But there is a natural reason why f^* above is called the Bayes classifer. Instead of taking $P \in \mathcal{P}$ to be the state of nature, we will consider $Y \in \{-1, 1\}$ to be the random state of nature, and set the action space as $\mathcal{A} = \{-1, 1\}$. The observations $X \in \mathcal{X}$, given the state Y, are then distributed as $P_{X|Y}(\cdot|Y)$. The decision rule $\delta : \mathcal{X} \mapsto \{-1, +1\}$ is then precisely a binary classifier that maps the inputs in $\mathcal{X}$ to outputs in $\{-1, +1\}$. The decision theoretic loss is then set to $L(Y, a) = \mathbb{I}[Y \neq a]$. The corresponding risk for a decision rule $\delta(\cdot)$ is then given as:

$$\begin{aligned} R(Y, \delta) &= \mathbb{E}_{X \sim P(\cdot|Y)} L(Y, \delta(X)) \\ &= \mathbb{E}_{X \sim P(\cdot|Y)} \mathbb{I}(Y \neq \delta(X)) \end{aligned}$$

so that with a prior distribution $P_Y(\cdot)$ over the state Y, we get the Bayes risk:

$$\begin{aligned} r(P_Y, \delta) &= \mathbb{E}_{Y \sim P_Y} \mathbb{E}_{X \sim P_{X|Y}(\cdot|Y)} \mathbb{I}(Y \neq \delta(X) \\ &= \mathbb{E}_{(X,Y) \sim P} \mathbb{I}(Y \neq \delta(X), \quad \text{where } P(X, Y) = P_Y(Y) P_{X|Y}(X|Y), \end{aligned}$$

which is precisely the mis-classification error $L(P, \delta)$. Thus, the classifier minimizing the mis-classification error f^* above is precisely the Bayes optimal classifier under this decision theoretic setup. Unlike typical Bayesian estimation settings however, both the prior P_Y, as well as the observation distribution $P_{X|Y}$ is unknown here, so that this is not actually actionable, but rather provides a characterization of the ideal classifier (i.e. assuming knowledge of $P(X, Y)$).

We note that there exist many other popular loss functions for binary classification that are not decomposable; see Koyejo et al. [2014] for a study of such non-decomposable loss functions. For instance, consider the precison loss function which is the fraction of true positives to the total number of predicted positives, so that

$$L(P, f) = P(Y = 1 | f(X) = 1) = \frac{P(Y = 1, f(X) = 1)}{P(f(X) = 1)},$$

which is not of the form $\mathbb{E}_{(X,Y) \sim P} \ell(f, (X, Y))$ for any $\ell(\cdot)$.

6.2 Plugin estimators

A close cousin of empirical risk minimization based estimators are so-called plugin estimators. We can distinguish between two classes of plugin estimators. In the first, we compute a plugin estimate of the loss itself, so that we approximate $L(P, f)$ by $L(P_n, f)$, where P_n is

the empirical distribution given samples $\{X_i\}_{i=1}^n \sim P$. This then allows us to compute:

$$\widehat{f}_{\text{PLUGIN;I}} = \arg \inf_{f \in \mathcal{F}} L(P_n, f). \tag{1}$$

Note that this does not require that the loss function be decomposable. For instance, for the precision loss above, this would entail solving for:

$$\arg \inf_{f \in \mathcal{F}} \frac{\sum_{i=1}^n \mathbb{I}[f(X_i = Y_i = 1)]}{\sum_{i=1}^n \mathbb{I}[f(X_i = 1)]}.$$

For decomposable losses, it can be seen that Eqn (1) directly reduces to empirical risk minimizers (ERMs).

The second class of plugin estimators is to first characterize the ideal estimator $f^*(P) = \arg\inf_f L(P, f)$, and then directly compute the plugin estimate: $\widehat{f}_{\text{PLUGIN;II}} = f^*(P_n)$. This is not always a good idea in this exact form. For instance, with the zero-one loss function, we get $\widehat{f}_{\text{PLUGIN;II}} = \text{sign}(P_n(Y = 1|\cdot) - 1/2)$, where $P_n(Y = 1|X)$ is simply the empirical conditional distribution, but which for continuous inputs X, will thus likely reduce to random guessing. Much more common is a related variant of computing a smoothed variant $\widetilde{P}_n$ (or for instance, fitting some statistical model such as logistic regression), and then using the plugin estimate $f^*(\widetilde{P}_n)$.

7 Characterization of Decomposable Losses

A natural question then is: what classes of loss functions $L(\theta^*, a)$ are decomposable, and hence amenable to ERM like estimators? Note that in its general form, this loss could depend arbitrarily on the state of nature θ^*, and the action a. For instance, the ℓ_p loss: $L(\theta^*, a) = \|\theta^* - a\|_p$ does not seem decomposable at all.

Denote the distribution over the observations given the state of nature by P_{θ^*}, and fix the action a. Assuming that the observation distributions are identifiable (i.e. we can recover θ from P_θ), the loss $L(\theta^*, a)$ could then be viewed as a functional of the distribution P_{θ^*}. For the loss to be decomposable as defined earlier, it would entail that:

$$L(\theta^*, a) \equiv L_a(P_{\theta^*}) = \mathbb{E}_{Z \in P_{\theta^*}} \ell_a(Z),$$

where we have used the overloaded notation $\ell_a(Z) := \ell(a, Z)$.

Generalizing this requirement, we can ask the following general question: given any distribution P, what loss functionals $\mathcal{L}(P)$ can be expressed as the expectation of some auxiliary loss evaluated at a random variable with distribution P? In other words, when can we write:

$$\mathcal{L}(P) = \mathbb{E}_{Z \sim P}(\ell(Z)), \tag{2}$$

for some auxiliary loss function $\ell(\cdot)$ of a random variable with the same distribution as the argument to the loss function $\mathcal{L}(\cdot)$. As we saw earlier, not all possible loss functionals can have this form, but classical results from Utility Theory [Berry, 1982] state some very reasonable sufficient conditions under which any loss functional will necessarily have the above form.

Let $\mathcal{P}$ be some class of distributions, and some loss functional $\mathcal{L} : \mathcal{P} \mapsto \mathbb{R}$

Axiom A. If $\mathcal{L}(P_1) < \mathcal{L}(P_2)$, then $\mathcal{L}(\alpha P_1 + (1-\alpha)P_3) < \mathcal{L}(\alpha P_2 + (1-\alpha)P_3)$, for any $\alpha \in (0,1)$, and $P_3 \in \mathcal{P}$.

This states that if P_1 has lower loss than P_2, then given two random situations which are identical except that in one P_1 occurs with probability α, while in the other P_2 occurs with probability alpha; the first random situation has lower loss.

Axiom B. If $\mathcal{L}(P_1) < \mathcal{L}(P_2) < \mathcal{L}(P_3)$, there exist $\alpha, \beta \in (0,1)$ s.t.

$$\mathcal{L}(\alpha P_1 + (1-\alpha)P_3) < \mathcal{L}(P_2), \quad \text{and} \quad \mathcal{L}(P_2) < \mathcal{L}(\beta P_1 + (1-\beta)P_3).$$

This axiom loosely states that there are no infinitely bad or good distributions. A sufficient condition for this Axiom to hold is that the loss functional be bounded over $\mathcal{P}$.

Theorem 2 (Degroot, 76 (adapted to our loss functional setting)) *Suppose the loss functional $\mathcal{L} : \mathcal{P} \mapsto \mathbb{R}$, over distributions in some set of distributions $\mathcal{P}$, satisfies the two axioms above. Then, the loss functional has the form $\mathcal{L}(P) = \mathbb{E}_{Z\sim P}\ell(Z)$, for all $P \in \mathcal{P}$.*

Thus the class of decomposable loss functionals encompasses all "rational" loss functionals that satisfy the very reasonable axioms above.

7.1 Uniform Optimality Bounds via Uniform Laws, Generalization Bounds

Recall the goal of uniform optimality bounds. For the specific class of ERM estimators, there is a particular technical tool that helps us obtain such pointwise bounds, namely, so-called uniform laws. These provide uniform guarantees of the deviation of the empirical loss (risk in ML terminology) from the true loss (risk):

$$r_{n;\theta^*} := \sup_{a\in\mathcal{A}} |\widehat{L}_n(\theta^*, a) - L(\theta^*, a)|,$$

that hold with high probability. When we can uniformly bound these:

$$r_{n,\theta^*} \leq r_{n;\text{unif}},$$

for all $\theta^* \in \Theta$, we can thus make these "distribution-free" (so that they do not depend on the model θ).

Given such a uniform law bound, we could then provide guarantees on the empirical risk minimizer (ERM):

$$\widehat{a}_{\text{erm}} := \arg\min_{a \in \mathcal{A}} \widehat{L}_n(\theta^*, a),$$

by a simple chaining argument:

$$\begin{aligned} L(\theta^*, \widehat{a}_{\text{erm}}) - L(\theta^*, a^*) &\leq L(\theta^*, \widehat{a}_{\text{erm}}) - \widehat{L}_n(\theta^*, \widehat{a}_{\text{erm}}) \\ &+ \widehat{L}_n(\theta^*, \widehat{a}_{\text{erm}}) - \widehat{L}_n(\theta^*, a^*) \\ &+ \widehat{L}_n(\theta^*, a^*) - L(\theta^*, a^*) \\ &\leq 2r_{n;\text{unif}}, \end{aligned}$$

since the first and third terms are bounded by the uniform bound, and the second term is bounded by zero, since by construction, $\widehat{a}_{\text{erm}}$ is the minimizer of the empirical risk.

Another class of bounds are so-called **generalization bounds**, where we bound the difference between empirical risk and true risk (note that we use the ML terminology here, and mean the losses rather than the expectation of these over the dataset) for the ERM estimator specifically, so that:

$$L(\theta^*, \widehat{a}_{\text{erm}}) \leq \widehat{L}_n(\theta^*, \widehat{a}_{\text{erm}}) + r_{n;\theta^*;\text{gen}},$$

where it can be seen that $r_{n;\theta^*,\text{gen}} \leq r_{n;\theta^*,\text{unif}}$, so that these are always tighter.

References

James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

James O Berger and Robert L Wolpert. The likelihood principle. IMS, 1988.

Oluwasanmi Koyejo, Nagarajan Natarajan, Pradeep Ravikumar, and Inderjit S Dhillon. Consistent binary classification with generalized performance metrics. In *NIPS*, volume 27, pages 2744–2752. Citeseer, 2014.

Donald A Berry. Statistical decision theory, foundations, concepts, and methods, 1982.