Statistical Decision and Learning Theory

Robert Nowak

Electrical and Computer Engineering,
University of Wisconsin, Madison, USA
nowak@engr.wisc.edu

Abstract. This paper reviews and contrasts the basic elements of *statistical decision theory* [1–4] and *statistical learning theory* [5–7]. It is not intended to be a comprehensive treatment of either subject, but rather just enough to draw comparisons between the two.

Throughout this paper, let X denote the *input* to a decision-making process and Y denote the correct response or *output* (e.g., the value of a parameter, the label of a class, the signal of interest). We assume that X and Y are random variables or random vectors with joint distribution $P_{X,Y}(x, y)$, where x and y denote specific values that may be taken by the random variables X and Y, respectively. The observation X is used to make decisions pertaining to the quantity of interest. For the purposes of illustration, we will focus on the task of determining the value of the quantity of interest. A decision rule for this task is a function f that takes the observation X as input and outputs a prediction of the quantity Y. We denote a decision rule by $\widehat{Y}$ or $f(X)$, when we wish to indicate explicitly the dependence of the decision rule on the observation. We will examine techniques for designing decision rules and for analyzing their performance.

0.1 Measuring Decision Accuracy: Loss and Risk Functions

The accuracy of a decision is measured with a loss function. For example, if our goal is to determine the value of Y, then a loss function takes as inputs the true value Y and the predicted value (the decision) $\widehat{Y} = f(X)$ and outputs a non-negative real number (the "loss") reflective of the accuracy of the decision. Two of the most commonly encountered loss functions include:

1. 0/1 loss: $\ell_{0/1}(\widehat{Y}, Y) = \mathbf{I}_{\widehat{Y} \neq Y}$, which is the indicator function taking the value of 1 when $\widehat{Y} \neq Y$ and taking the value 0 when $\widehat{Y}(X) = Y$.
2. squared error loss: $\ell_2(\widehat{Y}, Y) = \|\widehat{Y} - Y\|_2^2$, which is simply the sum of squared differences between the elements of $\widehat{Y}$ and Y.

The 0/1 loss is commonly used in detection and classification problems, and the squared error loss is more appropriate for problems involving the estimation of a continuous parameter. Note that since the inputs to the loss function may be random variables, so is the loss.

A risk $R(f)$ is a function of the decision rule f, and is defined to be the expectation of a loss with respect to the joint distribution $P_{X,Y}(x,y)$. For example, the expected 0/1 loss produces the *probability of error* risk function; i.e., a simply calculation shows that $R_{0/1}(f) = E[(\mathbf{I}_{f(X)\neq Y}] = \Pr(f(X) \neq Y)$. The expected squared error loss produces the *mean squared error* MSE risk function, $R_2(f) = E[\|f(X) - Y\|_2^2]$.

Optimal decisions are obtained by choosing a decision rule f that minimizes the desired risk function. Given complete knowledge of the probability distributions involved (e.g., $P_{X,Y}(x,y)$) one can explicitly or numerically design an optimal decision rule, denoted f^*, that minimizes the risk function.

0.2 The Maximum Likelihood Principle

The conditional distribution of the observation X given the quantity of interest Y is denoted by $P_{X|Y}(x|y)$. The conditional distribution $P_{X|Y}(x|y)$ can be viewed as a generative model, probabilistically describing the observations resulting from a given value, y, of the quantity of interest. For example, if y is the value of a parameter, the $P_{X|Y}(x|y)$ is the probability distribution of the observation X when the parameter value is set to y. If X is a continuous random variable with conditional density $p_{X|Y}(x|y)$ or a discrete random variable with conditional probability mass function (pmf) $p_{X|Y}(x|y)$, then given a value y we can assess the probability of a particular measurment value y by the magnitude of either the conditional density or pmf.

In decision making problems, we know the value of the observation, but do not know the value y. Therefore, it is appealing to consider the conditional density or pmf as a function of the unknown values y, with X fixed at its observed value. The resulting function is called the likelihood function. As the name suggests, values of y where the likelihood function is largest are intuitively reasonable indicators of true value of the unknown quantity, which we will denote by y^*. The rationale for this is that these values would produce conditional densities or pmfs that place high probability on the observation $X = x$.

The Maximum Likelihood Estimator (MLE) is defined to be the value of y that maximizes the likelihood function; i.e., in the continuous case

$$\widehat{y}(X) \;=\; \arg\max_y p_{X|Y}(X|y)$$

with an analogous definition for the discrete case by replacing the conditional density with the conditional pmf. The decision rule $\widehat{y}(X)$ is called an "estimator," which is common in decision problems involving a continuous parameter. Note that maximizing the likelihood function is equivalent to minimizing the negative log-likelihood function (since the logarithm is a monotonic transformation). Now let y^* denote the true value of Y. Then we can view the negative log-likelihood as a loss function

$$\ell_L(y, y^*) \;=\; -\log p_{X|Y}(X|y)$$

where the dependence on y^* on the right hand side is embodied in the observation X on the left. An interesting special case of the MLE results when the condi-

tional density $P_{X|Y}(X|y)$ is a Gaussian, in which case the negative log-likelihood corresponds to a squared error loss function.

Now let us consider the expectation of this loss, with respect to the conditional distribution $P_{X|Y}(X|y^*)$:

$$-E[\log p_{X|Y}(X|y)] = \int \log\left(\frac{1}{p_{X|Y}(x|y)}\right) p_{X|Y}(x|y^*)dx$$

The true value y^* minimizes the expected negative log-likelihood (or, equivalently, maximizes the expected log-likelihood). To see this, compare the expected log-likelihood of y^* with that of any other value y:

$$\begin{aligned} E[\log p_{X|Y}(X|y^*) - \log p_{X|Y}(X|y)] &= E\left[\log\left(\frac{p_{X|Y}(X|y^*)}{p_{X|Y}(X|y)}\right)\right] \\ &= \int \log\left(\frac{p_{X|Y}(x|y^*)}{p_{X|Y}(x|y)}\right) p_{X|Y}(x|y^*)dx \\ &= \mathrm{KL}(p_{X|Y}(x|y^*), p_{X|Y}(x|y)) \end{aligned} \quad (1)$$

The quantity $\mathrm{KL}(p_{X|Y}(x|y^*), p_{X|Y}(x|y))$ is called the Kullback-Leibler (KL) divergence between the conditional density function $p_{X|Y}(x|y^*)$ and $p_{X|Y}(x|y)$. The KL divergence is non-negative, and zero if and only if the two densities are equal [4]. So, we see that the KL divergence acts as a sort of risk function in the context of Maximum Likelihood Estimation.

0.3 The Cramer-Rao Lower Bound

The MLE is based on finding the value for Y that maximizes the likelihood function. Intuitively, if the maximum point is very distinct, say a well isolated peak in the likelihood function, then the easier it will be to distinguish the MLE from alternative decisions. Consider the case in which Y is a scalar quantity. The "peakiness" of the log-likelihood function can be gauged by examining its curvature, $-\frac{\partial^2 \log p_{X|Y}(x|y)}{\partial y^2}$, at the point of maximum likelihood. The higher the curvature, the more peaky is the behavior of the likelihood function at the maximum point. Of course, we hope that the MLE will be a good predictor (decision) for the unknown true value y^*. So, rather than looking at the curvature of the log-likelihood function at the maximum likelihood point, a more appropriate measure of how easily it will be to distinguish y^* from the alternatives is the expected curvature of the log-likelihood function evaluated at the value y^*. The expectation taken over all possible observations with respect to the conditional density $p_{X|Y}(x|y^*)$. This quantity, denoted $I(y^*) = E[-\frac{\partial^2 \log p_{X|Y}(x|y)}{\partial y^2}]|_{y=y^*}$, is called the Fisher Information (FI). In fact, the FI provides us with an important performance bound known as the Cramer-Rao Lower Bound (CRLB).

The CRLB states that under some mild regularity assumptions about the conditional density function $p_{X|Y}(x|y)$, the variance of any unbiased estimator

is bounded from below by the inverse of the $I(y^*)$ [1–3]. Recall that an unbiased estimator is any estimator $\widehat{Y}$ that satisfies $E[\widehat{Y}] = y^*$. The CRLB tells us is that

$$\text{var}(\widehat{Y}) \;\geq\; \frac{1}{I(y^*)}$$

If Y is a vector-valued quantity, then the expected negative Hessian matrix (matrix of partial second derivatives) of the log-likelihood function is called the Fisher Information Matrix (FIM), and a similar inequality tells us that the variance of each component of any unbiased estimator of y^* is bounded below by the corresponding diagonal element of the inverse of the FIM. Since the MSE of an unbiased estimator is equal to its variance, we see that the CRLB provides a very useful lower bound on the best MSE performance that we can hope to achieve. Thus, the CRLB is often used as a comparison point for evaluating estimators. It may or may not be possible to achieve the CRLB, but if we find a decision rule that does, we know that it also minimizes the MSE risk among all possible unbiased estimators. In general, it may be difficult to compute the CRLB, but in certain important cases it is possible to find closed-form or computational solutions.

0.4 Bayesian Decision Theory

Bayesian Decision Theory provides a formal system for integrating prior knowledge and observed observations. For the purposes of illustration we will focus on problems involving continuous variables and observations, but extensions to discrete cases are straightforward (simple replace probability densities with probability mass functions, and integrals with summations). The key elements of Bayesian methods are:

1. a prior probability density function $p_Y(y)$ describing a priori knowledge of probable states for the quantity Y;
2. the likelihood function $p_{X|Y}(x|y)$, as described above;
3. the posterior density function $p_{Y|X}(y|x)$.

The posterior density is a function of the prior and likelihood, obtained according to Bayes rule:

$$p_{Y|X}(y|x) = \frac{p_{X|Y}(x|y)p_Y(y)}{\int p_{X|Y}(x|y)p_Y(y)dy}$$

The posterior is an indicator of probable values for Y, based on the prior knowledge and the observation. Several options exist for deriving a specific estimate of Y using the posterior. The mean value of the posterior density is one common choice (commonly called the *posterior mean*). The posterior mean is the decision rule that minimizes the expected squared error loss (MSE risk) function. The value y where the posterior density is maximized is another popular estimator (commonly called the *Maximum A Posteriori* (MAP) estimator). Note that the denominator of the posterior is independent of y, so the MAP estimator is simply the maximizer of the product of the likelihood and the prior. Therefore, if the prior is a constant function, the MAP estimator and MLE coincide.

0.5 Statistical Learning

In all of the methods described above, we assumed some amount of knowledge about the distributions of the observation X and quantity of interest Y. Such knowledge can come from a careful analysis of the physical characteristics of the problem at hand, or it can be gleaned from previous experience. However, there are situations where it is difficult to model the physics of the problem and we may not have enough experience to develop complete and accurate probability models. In such cases, it is natural to adopt a *statistical learning* approach [5, 6].

Statistical learning methods are based on developing decision rules or estimators based only on a collection of training examples, rather than predetermined probability models. Statistical learning methods are often said to be *distribution-free*, since they do not assume particular probability models. The canonical set-up for statistical learning is as follows. We begin with a collection of training examples, $\{(X_i, Y_i)\}_{i=1}^n$, which are assumed to be independently and identically distributed according to an *unknown* probability distribution $P_{X,Y}(x, y)$. If we knew $P_{X,Y}(x, y)$, then we could compute a desired risk function and design an optimal decision rule using the methods described above. In essence, the training examples give us a glimpse at the underlying distribution, but our knowledge of it is far from complete. We cannot exactly compute a risk function, and therefore we cannot derive a corresponding optimal decision rule.

There are at least two ways to proceed at this point. One possibility is to use the training examples to estimate the joint probability distribution, and then use this estimate to derive an decision rule. Unfortunately, the (general-purpose) problem of estimating a distribution is often more difficult from a limited pool of data than is the problem of designing a specific-purpose decision rule. For this reason, a second possibility is more commonly favored in practice. Rather than estimating the complete distribution, one can use the training examples to directly design a decision rule. More precisely, perhaps the most common approach is to use the training examples to compute an estimate of the desired risk function.

Suppose that we are interested in minimizing a particular risk function. Recall that the risk is the expected value of a chosen loss function. Let $\ell(\widehat{Y}, Y)$ denote the loss, and let $f(X)$ denote a candidate decision function, mapping observations to predictions about Y (i.e., $\widehat{Y} = f(X)$). The *empirical risk function* is constructed from the training examples as follows:

$$\widehat{R}(f) \;=\; \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i)$$

This is simply the average loss of the decision rule f over the set of training examples. Note that since the training examples are independent and identically distributed, the expected value of the empirical risk is equal to the true risk $R(f) = E[\ell(f(X), Y)]$. Moreover, we known (according to the law of large numbers) that the empirical risk tends to the true risk as the size of the training sample increases. These facts lend support to the idea of choosing a decision rule to minimize the empirical risk.

Empirical risk minimization (ERM) is just this process. Given a collection of possible decision rules, say $\mathcal{F}$, ERM selects a decision rule according to

$$\widehat{f}_n \;=\; \arg\min_{f\in\mathcal{F}} \widehat{R}(f)$$

The selected rule, $\widehat{f}_n$, obviously depends on the given set of training examples, and therefore it is itself a random quantity. The theoretically optimal counterpart to $\widehat{f}_n$ is the decision rule that minimizes the true risk

$$f^* \;=\; \arg\min_{f\in\mathcal{F}} R(f)$$

The central problem in statistical learning is to quantify how close $\widehat{f}_n$ performs relative to f^*. Note that $R(f^*) \leq R(\widehat{f}_n)$, since f^* minimizes the true risk. Thus, one way to gauge the performance of $\widehat{f}_n$ relative to f^* is to show that there exists small positive values ϵ and δ such that with probability at least $1-\delta$ we have

$$R(\widehat{f}_n) \;\leq\; R(f^*) + \epsilon$$

If an inequality of this form holds, then we say that $\widehat{f}_n$ is a *Probability Approximately Correct* (PAC) decision rule [7].

To show that the empirical risk minimizer is a PAC decision rule, we first must understand how closely the empirical risk matches the true risk. First, let us consider the empirical and true risk of the decision rule f. Assume that the loss function is bounded between 0 and 1 (possibly after a suitable normalization). Then the empirical risk function is a sum of independent random variables bounded between 0 and 1. Hoeffding's inequality is a bound on the deviations of such random sums from their corresponding mean values [5]. In this case, the mean value is the true risk of f, and Hoeffding's inequality states that

$$P(|\widehat{R}(f) - R(f)| > \epsilon) \;\leq\; 2e^{-2n\epsilon^2}$$

Another equivalent statement is that the inequality $|\widehat{R}(f) - R(f)| \leq \epsilon$ holds with probability at least $1 - 2e^{-2n\epsilon^2}$. Thus, the two risks are probably close together, and the greater the number of training examples, n, the closer they are.

Now we would like a similar condition to hold for all $f \in \mathcal{F}$, since ERM optimizes over the entire collection $\mathcal{F}$. Suppose that $\mathcal{F}$ is a finite collection of decision rules. Let $|\mathcal{F}|$ denote the number of rules in $\mathcal{F}$. The probability that difference between the true and empirical risks, of one or more of the decision rules, exceeds ϵ is bounded by the sum of the probabilities of each individual event of the form $|\widehat{R}(f) - R(f)| > \epsilon$, the so-called *Union of Events* bound. Therefore, with probability at least $1 - |\mathcal{F}|2e^{-2n\epsilon^2}$ we have that

$$|\widehat{R}(f) - R(f)| \leq \epsilon$$

for all $f \in \mathcal{F}$. Equivalently, setting $\delta = 2|\mathcal{F}|e^{-2n\epsilon^2}$, we have that with probability at least $1-\delta$ and for all $f \in \mathcal{F}$

$$|\widehat{R}(f) - R(f)| \leq \sqrt{\frac{\log|\mathcal{F}| + \log(2/\delta)}{2n}}$$

Notice that the two risks are uniformly close together, and the closeness indicated by the bound increases as n increases and decreases as the number of decision rules in $\mathcal{F}$ increases. In fact, the bound scales with $\log |\mathcal{F}|$, and so it is reasonable to interpret the logarithm of the number of decision rules under consideration as a measure of the *complexity* of the class.

Now using this bound, we can show that $\widehat{f}_n$ is a PAC decision rule as follows. Note that with probability at least $1 - \delta$

$$\begin{aligned} R(\widehat{f}_n) &\leq \widehat{R}(\widehat{f}_n) + \sqrt{\frac{\log |\mathcal{F}| + \log(2/\delta)}{2n}} \\ &\leq \widehat{R}(f^*) + \sqrt{\frac{\log |\mathcal{F}| + \log(2/\delta)}{2n}} \\ &\leq R(f^*) + 2\sqrt{\frac{\log |\mathcal{F}| + \log(2/\delta)}{2n}} \end{aligned}$$

where the first inequality follows since the true and empirical risks are close for all $f \in \mathcal{F}$, and in particular for $\widehat{f}_n$, the second inequality holds since by definition $\widehat{f}_n$ minimizes the empirical risk, and the third inequality holds again since the empirical risk is close to the true risk for all f, in this case for f^* in particular. So, we have shown that $\widehat{f}_n$ is PAC.

PAC bounds of this form can be extended in many directions, for example to infinitely large or uncountable classes of decision rules, but the basic ingredients of the theory are essentially like those demonstrated above. The bottom line is that empirical risk minimization is a reasonable approach, provided one has access to a sufficient number of training examples and the number, or more generally the complexity, of the class of decision rules under consideration is not too great.

0.6 Further reading

Excellent treatments of classical decision and estimation theory can be found in a number of textbooks [1–4]. For references on statistical learning theory, outstanding textbooks are also available [5–7] for further reading.

References

1. Trees, H.L.V.: Detection, Estimation, and Modulation Theory, Part I. Wiley, New York (1968)
2. Lehmann, E.L.: Theory of Point Estimation. Wiley, New York (1983)
3. Kay, S.M.: Fundamentals of Statistical Signal Processing. Prentice Hall (1993)
4. Cover, T., Thomas, J.A.: Elements of Information Theory. Wiley (1991)
5. Devroye, L., Györfi, L., Lugosi, G.: A Probabilistic Theory of Pattern Recognition. Spring, New York (1996)
6. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)
7. Valiant, L.G.: A theory of the learnable. Communications of the ACM **27** (1984) 1134–1142