Stat 928: Statistical Learning Theory **Lecture: 0**

Risk vs Risk: Some differences between Statistics and ML Terminology

Instructor: Sham Kakade

1 A Quick Note

Unfortunately, the machine learning uses of the word "risk" quite differently from it's traditional definition in statistics — I am not sure what the historical reasons for this are. This is quick note to clarify things (mostly for my students).

My feeling is that it best not to use the terminology "risk" in ML, unless it is used in a manner consistent with the standard definition.

At a high level, the statistical risk measure the quality of the learning algorithm/estimation procedure (where the expectation is taken with respect to the training set).

2 From Training Sets to Parameter Estimation

We observe data:

$$\mathcal{T} = z_1, \ldots z_n$$

from some distribution.

For example, each z_i could just be a real number, sample from an i.i.d. distribution and our goal might be to just estimate the mean.

The standard example in supervised learning is where $z_i = (x_i, y_i)$, where have a distribution over input/output pairs. Our goal may be to predict the Y give some X, e.g. to predict the conditional expectation $\mathbb{E}[Y|X_i]$ if Y is real.

Typically, in supervised learning, we are interested in some notion of the our prediction loss. For example, in regression, the average squared error for a function f is:

$$L_{\text{squared error}}(f) = \mathbb{E}(f(X) - Y)^2$$

where the expectation is with respect to a random X, Y pair. In the fixed design setting, our prediction loss for a linear predictor was:

$$L_{\text{squared error}}(w) = \frac{1}{n}\mathbb{E}\|Xw - Y\|^2 = \frac{1}{n}\sum_i \mathbb{E}(X_i \cdot w - Y_i)^2$$

where now X is fixed matrix and Y is a vector.

Let us a say a decision rule δ is a mapping from $\mathcal{T}$ to some space — this is our estimation procedure or our learning algorithm. The notion of risk in statistics measures the quality of the procedure, on average.

3 Risk: In statistics

Suppose we have a set of distributions $\{P_\theta : \theta \in \Theta\}$. Assume there exists a θ^* such that:

$$\mathcal{T} \sim P_{\theta^*}$$

Say our goal is to estimate θ^*, and $\delta(\mathcal{T})$ is our estimate $\hat{\theta}$.

Suppose we a loss function which measures the error in our parameter estimate:

$$L_{\text{param}}(\theta^*, \theta)$$

Note this is a loss between parameter estimates.

For example, in the fixed design regression setting, the standard choice is:

$$L_{\text{param}}(\theta^*, \theta) = \|\theta - \theta^*\|_\Sigma^2$$

(where $\Sigma = \frac{1}{n} X^\top X$).

Note that, for linear regression,

$$L_{\text{param}}(w^*, w) = L_{\text{prediction}}(w) - L_{\text{prediction}}(w^*)$$

e.g. this parameter loss is the "regret" of w.

The *risk* is defined with respect to the true θ^* and the decision rule δ. It is defined as:

$$Risk_{\text{statistica}}(\theta^*, \delta) = \mathbb{E}_{\mathcal{T} \sim P_{\theta^*}} L_{\text{param}}(\theta^*, \theta(\mathcal{T}))$$

Critically, note that the expectation is over the training set.

4 Risk in Machine Learning

In the supervised learning setting, we have loss function $\ell(f(X), Y)$ which is the prediction loss on a given (X, Y) pair with respect to the function f. For example,

$$\ell(f(X), Y) = (f(X) - Y)^2$$

for regression.

Often, in machine learning, the "risk" is defined as:

$$Risk_{\text{ML}}(f) = \mathbb{E}[\ell(f(X), Y)]$$

where the expectation is with respect to the underlying distribution. Sometimes this is also referred to as the average loss, or the average prediction loss.

Often, this will be denoted as:

$$L(f) = \mathbb{E}[\ell(f(X), Y)]$$

(which will be the convention in this class). We will just refer to this as the average loss.

Let f^* is the minimizer of L in some set $\mathcal{F}$, e.g.

$$f^* \in \arg\min_{f \in \mathcal{F}} L(f)$$

The *regret* of f is defined as:

$$L(f) - L(f^*)$$

Note that for the case of square error:

$$L_{\text{param}}(w^*, w) = L(w) - L(w^*)$$

where the parameter loss is defined with respect to the appropriate norm between w^* and w. Essentially, the regret is often analogous to the parameter loss. Though often ML does not assume a true model for the data generation process (beyond iid and certain other restrictions).

4.1 Risk vs Risk

Again, let δ be the learning rule (e.g. a decision rule), which takes $\mathcal{T}$ to some function $f \in \mathcal{F}$. We are typically interested in statements of the form:

With probability greater than $1 - \delta$*,*

$$L(\delta(\mathcal{T})) - L(f^*) \leq ??$$

where the randomness is with respect to the training set $\mathcal{T}$*.*

The analogue of the statistical risk, in this setting, would be:

$$Risk_{\text{Better Definition}}(f^*, \delta) = \mathbb{E}_{\mathcal{T}}[L(\delta(\mathcal{T})) - L(f^*)]$$

and this definition is consistent in the regression setting. The important high level distinction is that, in statistics, the risk is a measure of the quality of the decision procedure (e.g. the learning algorithm).

{ The most natural analogy for the statistical risk in the machine learning terminology would be the expected risk of a learning algorithm, where the expectation is with respect to the training set. }

ML typically does not directly look at statements of the form:

$$\mathbb{E}_{\mathcal{T}}[L(\delta(\mathcal{T})) - L(f^*)] \leq ??$$

but instead attempts to make high probability statements. These high probability statements are essentially a stronger guarantees. Typically, bound the above expectation, usually involves going through a high probability argument anyways. Hence, often just the high probability statements are provided.

However, bounding even the expected regret (e.g. statistical risk) of the learning algorithm is powerful statement. Any strong guarantee on performance of the learning algorithm/decision procedure is ultimately what we are after.