Risk functions for multivariate prediction

Charles Zheng and Yuval Benjamini

October 14, 2015

Broadly speaking, the goal of *supervised learning* is to learn the conditional distribution of the response Y conditional on predictors x. Here we are interested in the case of where both the predictors $x \in \mathbb{R}^p$ and response $Y \in \mathbb{R}^q$ are high-dimensional. Later we will be particularly interested in the special case

$$Y|x \sim N(B^T x, \Sigma)$$

where the unknown parameters are B, a $p \times q$ coefficient matrix and Σ, a $q \times q$ covariance matrix.

But let us return now to the general case. Suppose that in truth, $Y|x$ has a distribution F_x. Based on training data, we estimate the distribution $Y|x$ as $\hat{F}_x$. Is $\hat{F}_x$ a good estimate of the truth, F_x? Well, it depends on what our ultimate goal is. If our goal is simply to produce a prediction $\hat{Y}$ that minimizes the squared error loss with the observed Y, then we should choose $\hat{Y} = \mathbf{E}_{\hat{F}_x} Y$, and hence the risk function we should use to evaluate our procedure is the usual squared-error prediction risk,

$$\text{risk}_{pred}(\hat{F}_x) = \mathbf{E}[||Y - \hat{Y}||^2] = \mathbf{E}[||Y - \mathbf{E}_{\hat{F}_x} Y||^2].$$

Supposing the covariate is also a random variable, then we want to average the above risk function over the random distribution of X, defining

$$\text{Risk}_{pred}(\hat{F}_X) = \mathbf{E}[\text{risk}_{pred}(\hat{F}_x)|X = x].$$

Yet, risk_{pred} is not the only risk function one could use. Assuming that F_x has a density f_x relative to some measure μ, one could define the Kullback-Liebler risk as

$$\text{risk}_{KL}(\hat{F}_x) = -\mathbf{E}[\log \hat{f}_x(Y)]$$

Unlike risk_{pred}, the Kullback-Liebler loss requires us to get a good estimate of the whole distribution, not just its mean. And as before, if X is random, we can define $\text{Risk}_{KL}(\hat{F}_X)$ similarly to before.

It could be expected that using different risk functions leads to different theoretical approaches and procedures. While risk_{pred} is one of the simpler cases, it already lends itself to sophisticated approaches involving simultaneous estimation of B and Σ: see, for instance Witten and Tibshirani (2008). Presumably, minimizing risk_{KL} would have to involve even more complicated procedures, if the problem is even tractable at the moment. Yet, researchers are often interested in knowing more than the conditional mean: hence it would be interesting to look at risk functions which are somewhat more involved than risk_{pred}, but which may be easier from both a theoretical and practical perspective than risk_{KL}. Note that both risk_{pred} and risk_{KL} have the property that they are minimized by the true value F_x:

$$\min \text{risk}(\hat{F}_x) = \text{risk}(F_x)$$

We might call a risk function "unbiased" if it has this property: not to be confused with the unbiasedness of the estimators! A unbiased risk function might still be minimized by a biased estimator. On the other hand, we can't imagine why one would ever want to study a biased risk function.

Stopping short of estimating the conditional distribution, one might evaluate the first two moments of $\hat{F}_x$, by using a loss function involving a term like

$$(Y - \hat{Y})^T \hat{\Sigma}^{-1} (Y - \hat{Y})$$

where $\hat{Y}$ is the mean of $\hat{F}_x$ and $\hat{\Sigma}$ is the covariance of $\hat{F}_x$. However, the above expression by itself does not represent an unbiased risk function, since it is minimized by $\hat{\Sigma} = \infty$ irrespective of the true distribution.

1.

2.