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 some map $\hat{F} : x \mapsto \hat{F}_x$, where $\hat{F}_x$ is an estimate of the distribution $Y|x$. 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

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

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 $\text{Risk}_{KL}(\hat{F})$ 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). Are there other interesting risk functions? 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 estimators! Biased estimators are quite useful, but it is hard to see why one would ever study a biased risk function.

The *covariance risk* which we are about to define is a familiar expression, but perhaps not outside of a Gaussian context:

$$\text{risk}_{\Sigma} = \mathbf{E}[(Y - \hat{Y})^T \hat{\Sigma}^{-1} (Y - \hat{Y})] + \log\det\hat{\Sigma}$$

where $\hat{Y}$ is the mean of $\hat{F}_x$ and $\hat{\Sigma}$ is the covariance of $\hat{F}_x$. It is easy to show that *covariance risk* risk_{Σ} is unbiased: first note that fixing $\hat{\Sigma}$, the risk is minimized by $\hat{Y} = \mathbf{E}[Y|x]$; using this choice of $\hat{Y}$, the risk simplifies to $\mathbf{E}\text{tr}(\Sigma\hat{\Sigma}^{-1}) + \log\det\hat{\Sigma}$, which is stationary at $\hat{\Sigma} = \Sigma$.

It is worth noting that prediction risk and covariance risk are equivalent to risk_{KL} for the gaussian model with identity covariance and unknown covariance, respectively.

However, there are yet more ways to evaluate the quality of the estimated map $\hat{F}$. Letting (X^*, Y^*) be a new covariate and response, I could give you the response Y^*, and ask you to predict the covariate X^*. Your goal would be to minimize the squared error risk for X^*:

$$\text{risk} = \mathbf{E}[||X^* - \hat{X}||^2]$$

Supposing you knew the distribution of X^* in advance, $p(x)$, you should choose

$$\hat{X} = \frac{\int x\hat{f}_x(Y^*)p(x)dx}{\int \hat{f}_x(Y^*)p(x)dx}.$$

In the context of neuroscience, the problem of predicting x given y, where x represents a stimulus and y a neuronal response, is known as *stimulus reconstruction.* Hence we define

$$\text{Risk}_{recon}(\hat{F}) = \mathbf{E}[||X^* - \hat{X}||^2]$$

where $\hat{X}$ is defined as $\frac{\int x\hat{f}_x(Y^*)p(x)dx}{\int \hat{f}_x(Y^*)p(x)dx}$.

But now suppose that X^* is initially sampled from a finite sample from $p(x)$: $x_1, \ldots, x_\ell$. If I gave you Y^* and also the initial *candidate set* $x_1, \ldots, x_\ell$, then there is a nonzero probability of being able to predict X^* exactly, and one uses 0-1 loss to evaluate the *misclassification risk*,

$$\text{risk} = \Pr[X^* \neq \hat{X}]$$

The advatantage here is that besides the choice of $x_1, \ldots, x_\ell$ (which can be randomized), the misclassification risk is in some sense less arbitrary than the reconstruction risk as it does not depend on the scaling of x. Under this risk, you ought to choose X^* by

$$\hat{X} = \text{argmax}_{x_1,\ldots,x_\ell} \hat{f}_x(y)$$

In the context of neuroscience, this classification task is known as *identification.* With $\hat{X}$ defined as above, we thus define

$$\text{Risk}_{ident,\ell}(\hat{F}) = \mathbf{E}_{x_1,\ldots,x_\ell \sim p(x)} \Pr[X^* \neq \hat{X}]$$

making it clear that the risk is averaged over the random draws of $x_1, \ldots, x_\ell$.

Finally, we have an expanded list of evaluation criteria:

1. Prediction risk Risk_{pred}
2. KL risk Risk_{KL}
3. Covariance risk Risk_{Σ}
4. Reconstruction risk Risk_{recon}
5. Identification risk Risk_{ident}

How does the choice of risk function affect the difficulty of the resulting supervised learning task? Can the same approaches be used for multiple problems in the list, or does each choice of risk function demand a particular set of approaches? In any case, a first practical step would be to look at these problems in the Gaussian linear model.

Under the Gaussian linear model, there appears to a connection between prediction under the covariance risk and the goal of constructing Frequentist prediction sets. Under the gaussian model, the prediction set takes the form of an ellipse C of the form

$$C = \{y : (y - \mu)^T \Sigma^{-1} (y - \mu) < t\}$$

for some threshold t. Obtaining *minimal volume* prediction set with a given coverage amounts to choosing (μ, Σ) so that the large deviations of $(y - \mu)^T \Sigma^{-1} (y - \mu)$ are controlled while the log volume, $\log\det\Sigma$ is not too large– notice how both terms appear in risk_Σ.

Based on this connection noted between prediction under covariance risk and construction of prediction sets, we speculate that for optimal prediction under covariance risk, the conditional covariance $\hat{\Sigma}|x$ should vary depending on x, as it does when constructing prediction sets.

We also speculate on a connection between prediction under covariance risk and prediction under identification risk. Due to gaussianity, the identification risk depends solely on the estimated conditional mean and covariance: one chooses X^* from $x_1, ..., x_\ell$ by

$$\text{argmin}_{x_i} (y - x^T \hat{B})^T \hat{\Sigma}_x^{-1} (y - x^T \hat{B}) + \log\det\hat{\Sigma}_x$$

which is exactly the covariance risk, but replacing X^* with x. The point is that one makes a mistake if the above expression is too large for $x = X^*$, hence one might expect to minimize mistakes by minimizing the covariance risk. Our claim is that any method of estimating $\mu|x$ and $\Sigma|x$ which is "approximately optimal" for prediction under covariance risk should also be "approximately optimal" optimal for identification (where "approximately optimal" remains to be defined!)