STAT675 – Homework 1
Due: Sept. 17

1. a. Show that the prediction (also known as generalization) squared-error risk can be written as
$$R(f) = \mathbb{E}_{X,Y}(f(X) - Y)^2 = \mathbb{E}_X(f(X) - \mathbb{E}[Y|X])^2 + \mathbb{E}_X[\mathbb{V}[Y|X]]. \tag{1}$$

   b. What does this imply about the Bayes rule for squared error loss?

2. Reminder from lecture: assume that we get a new draw of the training data, $\mathcal{D}^0$, such that $\mathcal{D} \sim \mathcal{D}^0$ and

$$\mathcal{D} = ((X_1, Y_1), \ldots, (X_n, Y_n)) \quad \text{and} \quad \mathcal{D}^0 = ((X_1, Y_1^0), \ldots, (X_n, Y_n^0))$$

If we make a small compromise to risk, we can form a sensible suite of risk estimators
To wit, letting $Y^0 = (Y_1^0, \ldots, Y_n^0)^\top$, define

$$R_{in} = \mathbb{E}_{Y^0|\mathcal{D}} \hat{\mathbb{P}}_{\mathcal{D}^0} \ell_{\hat{f}} = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{Y^0|\mathcal{D}} \ell(\hat{f}(X_i), Y_i^0).$$

Then the average optimism is

$$\text{opt} = \mathbb{E}_Y [R_{in} - \hat{R}_{\text{train}}] = \frac{2}{n} \sum_{i=1}^n \text{Cov}(\hat{f}(X_i), Y_i).$$

Therefore, we get the following estimate of risk

$$\mathbb{E}_Y R_{in} = \mathbb{E}_Y \hat{R}_{\text{train}} + \frac{2}{n} \sum_{i=1}^n \text{Cov}(\hat{f}(X_i), Y_i),$$

which has unbiased estimator (i.e. $\mathbb{E}_Y R_{\text{gic}} = \mathbb{E}_Y R_{in}$)

$$R_{\text{gic}} = \hat{R}_{\text{train}} + \frac{2}{n} \sum_{i=1}^n \text{Cov}(\hat{f}(X_i), Y_i).$$

Our task now is to either estimate or compute opt to produce $\widehat{\text{opt}}$ and form

$$\hat{R}_{\text{gic}} = \hat{R}_{\text{train}} + \widehat{\text{opt}}. \tag{2}$$

a. **Stein's lemma:**

   i. Let $Z \sim N(0,1)$ and let $f : \mathbb{R} \to \mathbb{R}$ be absolutely continuous with derivative f'. Then[1]

   $$\mathbb{E}[Zf(Z)] = \mathbb{E}[f'(Z)]$$

   Show this is true. See [6] for more details.

   ii. Extend this result to cover an arbitrary normal random variable $X \sim N(\mu, \sigma^2)$.

   iii. Suppose[2] $Y \sim (\mu, \sigma^2 I) \in \mathbb{R}^n$ and let $f : \mathbb{R}^n \to \mathbb{R}^n$. Show that the expected training error can be decomposed as

   $$\mathbb{E}||\mu - f(y)||_2^2 = -n\sigma^2 + \mathbb{E}||y - f(y)||_2^2 + 2\sum_{i=1}^n Cov(Y_i, f_i(Y)).$$

[1] Note: we may not return to this, but it turns out this is an if and only if statement
[2] This notation means Y has mean μ and variance $\sigma^2 I$.

iv. It is possible to show that for each $i = 1, \ldots, n$, as long as f_i is almost differentiable, then if $X \sim N(\mu, \sigma^2 I)$,

$$\frac{1}{\sigma^2}\mathbb{E}[(X-\mu)f_i(X)] = \mathbb{E}[\nabla f_i(X)],$$

where $\nabla f_i(X)$ is the gradient of the i^{th} component of f evaluated at X. Use this fact (which is a multivariate extension of i.) to get an unbiased estimator of the risk. This is known as Stein's Unbiased Risk Estimator (SURE). It is a generalization of Mallow's Cp. Note that $\sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x)$ is known as the divergence of f.

b. **Stein's paradox.** We will use Stein's lemma to show that the usual maximum likelihood estimator X for estimating μ in $X \sim N(\mu, \sigma^2 I) \in \mathbb{R}^n$ is inadmissible[3] when $n \geq 3$. It turns out that

$$\hat{\mu} = \left(1 - \frac{(d-2)\sigma^2}{||X||_2^2}\right) X$$

uniformly dominates X. See [5] for the original paper and [1] for a nontechnical discussion of this point.

i. What is the risk of X as an estimator of μ?

ii. Use your result from the previous question to compute the SURE of $\hat{\mu}$. Note: this will reduce to computing the training error and then the divergence of the estimator.

iii. Take the expectation of the SURE for $\hat{\mu}$ and show that its risk is always lower than that of X. Jensen's inequality will come in handy. Also, a result[4] about χ^2 random variables: suppose that W is a non-central $\chi^2_{\nu,\delta}$ random variable with non-centrality parameter δ and ν degrees of freedom. Then $W \sim \chi^2_{\nu+2K,0}$, where $K \sim Pois(\delta/2)$.

c. **Degrees of freedom.** Inline with the definitions above, let $Y_1, \ldots, Y_n$ be such that $\mathbb{V}Y_i = \sigma^2$ and $Cov(Y_i, Y_{i'}) = \sigma^2 \delta_{i,i'}$ (the Kronecker delta function). Let $g : \mathbb{R}^n \to \mathbb{R}^n$ be a function that gives be fitted values, ie: $g(Y_1, \ldots, Y_n) = \hat{Y} \in \mathbb{R}^n$. Then

$$\text{df}(g) = \frac{1}{\sigma^2}\sum_{i=1}^n Cov(Y_i, g_i(Y)) = \frac{1}{\sigma^2}\text{trace}(Cov(Y, g(Y))).$$

Therefore, we can use our results from the previous sections to calculate degrees of freedom for various fitting procedures. Let's do that for

i. Ridge regression

ii. For lasso, I don't want you to derive the degrees of freedom. Instead, look over [7] and see if you can following the general flow of the argument, at least up to the end of section 2.1. Give an overview of the argument here.

d. **Generalized information criterion (GIC).** The original proposed GIC was in [3] and had the following form. Assume $Y_i = X_i^\top \beta_* + \epsilon_i$, where $\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$. The main goal was model selection, so let $\alpha \in A = \{\text{candidate models}\}$, where this could be all $2^p - 1$ models from p covariates for instance. Then

$$\text{GIC}_0(\alpha) = \log(\hat{\sigma}^2_\alpha) + \frac{1}{n}\kappa_n d_\alpha,$$

where $\hat{\sigma}^2_\alpha$ is the MLE under model α, (κ_n) is a sequence of numbers, and d_α is the degrees of freedom from model α. Choosing $\kappa_n = 2$ produces AIC, $\kappa = \log(n)$ produces BIC.

[3] I'm going to leave it up to you to look up what inadmissible means

[4] Known as 'Poissonization'.

i. These choices work when $n >> p$. However, when $n \leq p$, this doesn't work at all. Why?

ii. Instead, we use equation (2), with $\widehat{\text{opt}} = \hat{\sigma}^2 \kappa_n d_\alpha / n$ and $\hat{\sigma}^2$ is an estimator of the variance (see [8]) for more information). How could you make this approach operational in practice?

References

[1] Bradley Efron and Carl N Morris. *Stein's paradox in statistics.* ., 1977.

[2] Yang Feng and Yi Yu. Consistent cross-validation for tuning parameter selection in high-dimensional variable selection. *arXiv preprint arXiv:1308.5390*, 2013.

[3] Ryuei Nishii. Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, 12(2):758–765, 1984.

[4] Stephen Reid, Robert Tibshirani, and Jerome Friedman. A study of error variance estimation in lasso regression. *arXiv preprint arXiv:1311.5274*, 2013.

[5] Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In University of California Press, editor, *Proceedings of the Third Berkeley Symposium on Mathematical Statististics and Probability*, volume 1, pages 197–206, 1956.

[6] Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151, 1981.

[7] Ryan J Tibshirani and Jonathan Taylor. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232, 2012.

[8] Yiyun Zhang, Runze Li, and Chih-Ling Tsai. Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105(489):312–323, 2010.