STAT675 – Homework 1 (solutions)
Due: Sept. 11

1. a. Show that the prediction (also known as generalization) squared-error risk can be written as

$$R(f) = \mathbb{E}_{X,Y}(f(X) - Y)^2 = \mathbb{E}_X(f(X) - \mathbb{E}[Y|X])^2 + \mathbb{E}_X[\mathbb{V}[Y|X]]. \tag{1}$$

Solution [Aaron][1]**:**

$$\begin{aligned} R(f) &= \mathbb{E}_{X,Y}(f(X) - Y)^2 \\ &= \mathbb{E}_X \mathbb{E}_{Y|X}(f(X) - Y)^2 \\ &= \mathbb{E}_X[f(X)^2 - 2f(X)E[Y|X] + E[Y^2|X]] \\ &= \mathbb{E}_X[f(X)^2 - 2f(X)E[Y|X] + \mathbb{V}[Y|X] + E[Y|X]^2] \\ &= \mathbb{E}_X[f(X)^2 - 2f(X)E[Y|X] + E[Y|X]^2] + \mathbb{E}_X[\mathbb{V}[Y|X]] \\ &= \mathbb{E}_X[f(X) - E[Y|X]]^2 + \mathbb{E}_X[\mathbb{V}[Y|X]] \end{aligned}$$

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

Solution [Aaron]:
The Bayes Rule with respect to the squared error loss function is the posterior mean. This can be seen as setting $f_*(x) = E[Y|X]$ minimizes the risk function.

[1] Names are given so that questions may be addressed to the proper person.

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[2]

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

Show this is true. See [**?**] for more details.

[2] Note: we may not return to this, but it turns out this is an if and only if statement

Solution [Aaron]:
Let $\phi(z)$ be a standard normal density.
Note: $z\phi(z) = -\phi(z)$

$$
\begin{aligned}
E[f'(Z)] &= \int_{-\infty}^{\infty} f'(z)\phi(z)dz \\
&= \int_{0}^{\infty} f'(z)\phi(z)dz + \int_{-\infty}^{0} f'(z)\phi(z)dz \\
&= \int_{0}^{\infty} f'(z)\int_{z}^{\infty} y\phi(y)dydz - \int_{-\infty}^{0} f'(z)\int_{-\infty}^{z} y\phi(y)dydz \\
&= \int_{0}^{\infty}\int_{z}^{\infty} f'(z)y\phi(y)dydz - \int_{-\infty}^{0}\int_{-\infty}^{z} f'(z)y\phi(y)dydz \\
&= \int_{0}^{\infty}\int_{0}^{y} f'(z)y\phi(y)dzdy - \int_{-\infty}^{0}\int_{y}^{\infty} f'(z)y\phi(y)dzdy \\
&= \int_{0}^{\infty} [f(y) - f(0)]y\phi(y)dy + \int_{-\infty}^{0} [f(y) - f(0)]y\phi(y)dy \\
&= \int_{-\infty}^{\infty} [f(y) - f(0)]y\phi(y)dy \\
&= \mathbb{E}[Z(f(Z) - f(0))] \\
&= \mathbb{E}[Zf(Z)]
\end{aligned}
$$

Integration can be interchanged in the above derivation by Fubini's Theorem and the absolute continuity of f

ii. Extend this result to cover an arbitrary normal random variable $X \sim N(\mu, \sigma^2)$.
Solution [Aaron]:
Define a new function $h : \mathbb{R} \to \mathbb{R}$ by $h(x) = g(\frac{x-\mu}{\sigma})$

$$
\begin{aligned}
E[h'(X)] &= \frac{1}{\sigma}\mathbb{E}[f'(\frac{X-\mu}{\sigma})] \\
&= \frac{1}{\sigma}\mathbb{E}[f'(Z)] \\
&= \frac{1}{\sigma}\mathbb{E}[Zf(Z)] \\
&= \frac{1}{\sigma}\mathbb{E}[(\frac{X-\mu}{\sigma})f(\frac{X-\mu}{\sigma})] \\
&= \mathbb{E}[(\frac{X-\mu}{\sigma^2})h(X)]
\end{aligned}
$$

iii. Suppose[3] $Y \sim (\mu, \sigma^2 I) \in \mathbb{R}^n$ and let $f : \mathbb{R}^n \to \mathbb{R}^n$. Show that the expected training error can be decomposed as

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

[3] This notation means Y has mean μ and variance $\sigma^2 I$.

Solution [Ben]:

$$\begin{aligned}\mathbb{E}||\mu - f(y)||_2^2 &= E||\mu - y||_2^2 - 2E < \mu - y, f(y) - y > + \mathbb{E}||f(y) - y||_2^2 \\ &= (I) + (II) + (III)\end{aligned}$$

$$\begin{aligned}(I) &= \mathbb{E}[\mu - y]^T \mathbb{E}[\mu - y] + tr(cov(\mu - y)) \\ &= tr(cov(\mu - y)) \\ &= n\sigma^2\end{aligned}$$

$$\begin{aligned}(II) &= -2(\mathbb{E} < \mu - y, f(y) - \mathbb{E}[f(y)] > + \mathbb{E} < \mu - y, \mathbb{E}[f(y)] - y >) \\ &= 2\sum_{i=1}^{n} Cov(Y_i, f_i(Y)) - 2(\mathbb{E}[(\mu - y)]^T \mathbb{E}[f(y)] + \mathbb{E}[< \mu - y, y >]) \\ &= 2\sum_{i=1}^{n} Cov(Y_i, f_i(Y)) - 2\mathbb{E}[< \mu - y, y >] \\ &= 2\sum_{i=1}^{n} Cov(Y_i, f_i(Y)) + 2\mu^T\mu - 2\mathbb{E}[y^T y] \\ &= 2\sum_{i=1}^{n} Cov(Y_i, f_i(Y)) + 2\mu^T\mu - 2(\mu^T\mu + n\sigma^2) \\ &= 2\sum_{i=1}^{n} Cov(Y_i, f_i(Y)) - 2n\sigma^2\end{aligned}$$

Summing (I), (II), and (III) yields the desired decomposition

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.

Solution [Ben]:

First, we note that

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

implies that

$$\frac{1}{\sigma^2} cov(X, f(X)) = \mathbb{E}[\nabla f_1(X) \nabla f_2(X) ... \nabla f_n(X)]. \tag{3}$$

Now, from 2.a.iii, we have that

$$\frac{trace(cov(X,f(X)))}{\sigma^2} = \frac{Risk + n\sigma^2 - \mathbb{E}||X - f(X)||_2^2}{2\sigma^2},$$

Thus, from (3),

$$trace(\mathbb{E}[\nabla f_1(X)\nabla f_2(X)...\nabla f_n(X)]) = \frac{Risk + n\sigma^2 - \mathbb{E}||X - f(X)||_2^2}{2\sigma^2}.$$

Rearranging this equality, we see that

$$\begin{aligned} Risk &= 2\sigma^2 trace(\mathbb{E}[\nabla f_1(X)\nabla f_2(X)...\nabla f_n(X)]) - n\sigma^2 + \mathbb{E}||X - f(X)||_2^2 \\ &= 2\sigma^2 \mathbb{E}[trace([\nabla f_1(X)\nabla f_2(X)...\nabla f_n(X)]) - n\sigma^2 + \mathbb{E}||X - f(X)||_2^2 \\ &= \mathbb{E}[2\sigma^2 \sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i} - n\sigma^2 + ||X - f(X)||_2^2] \end{aligned}$$

.

Thus,

$$2\sigma^2 \sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i} - n\sigma^2 + ||X - f(X)||_2^2$$

is an unbiased estimator of the risk.

b. **Stein's paradox.** We will use Stein's lemma to show that the usual maximum likelihood estimator X for estimating μ in $X \sim \mathcal{N}(\mu, \sigma^2 I) \in \mathbb{R}^n$ is inadmissible[4] 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 [**?**] for the original paper and [**?**] for a nontechnical discussion of this point.

Solution (i. - iii.) [Henry]:

i. The risk of X as an estimator of μ is

$$\begin{aligned} R(X) &= \mathbb{E}||X - \mu||_2^2 \\ &= \sum_{i=1}^{d} \mathbb{E}(X_i - \mu_i)^2 \\ &= d\sigma^2. \end{aligned}$$

ii. The SURE of $\hat{\mu}$ involves the training error

$$\begin{aligned} ||X - \hat{\mu}||_2^2 &= ||X - \left(1 - \frac{(d-2)\sigma^2}{||X||_2^2}\right) X||_2^2 \\ &= ||\frac{(d-2)\sigma^2}{||X||_2^2} X||_2^2 \\ &= \frac{(d-2)^2\sigma^4}{||X||_2^4}||X||_2^2 \\ &= \frac{(d-2)^2\sigma^4}{||X||_2^2} \end{aligned}$$

[4] A numerical exploration of how this and how it applies to two other distributions under a certain class of loss functions can be had here: https://fairyaunts.shinyapps.io/steins_paradox/.

and the divergence of the estimator

$$\begin{aligned}
\nabla\hat{\mu} &= \sum_{i=1}^{d} \frac{\partial}{\partial x_i}\left(1 - \frac{(d-2)\sigma^2}{||X||_2^2}\right) x_i \\
&= \sum_{i=1}^{d} \left(1 - \frac{(d-2)\sigma^2}{||X||_2^2}\right) + \frac{(d-2)\sigma^2}{||X||_2^4}(-1)(2x_i)x_i \\
&= d - \frac{d(d-2)\sigma^2}{||X||_2^2} + 2\frac{(d-2)\sigma^2||X||_2^2}{||X||_2^4} \\
&= d - \frac{(d-2)^2\sigma^2}{||X||_2^2}.
\end{aligned}$$

Using the result from 2a.iv. gives

$$\begin{aligned}
SURE(\hat{\mu}) &= -d\sigma^2 + \frac{(d-2)^2\sigma^4}{||X||_2^2} + 2d\sigma^2 - 2\frac{(d-2)^2\sigma^4}{||X||_2^2} \\
&= d\sigma^2 - \frac{(d-2)^2\sigma^4}{||X||_2^2}.
\end{aligned}$$

iii. The expected value of this risk estimate is

$$\mathbb{E}\left[SURE(\hat{\mu})\right] = d\sigma^2 - (d-2)^2\sigma^4\mathbb{E}\left[\frac{1}{||X||_2^2}\right]$$

which we can bound from above using Jensen's inequality:

$$\begin{aligned}
&\leq d\sigma^2 - (d-2)^2\sigma^4\frac{1}{\mathbb{E}||X||_2^2} \\
&= d\sigma^2 - \frac{(d-2)^2\sigma^2}{d}
\end{aligned}$$

For all $d > 2, \sigma > 0$ the second term will be strictly positive, and so

$$\mathbb{E}\left[SURE(\hat{\mu})\right] < d\sigma^2 = \mathbb{E}\left[X\right]$$

for all μ, which proves the inadmissibility of X as an estimator for μ.

c. **Degrees of freedom.** Inline with the definitions above, let $Y_1, \dots, Y_n$ 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: $g(Y_1, \dots, Y_n) = \hat{Y} \in \mathbb{R}^n$. 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
Solution [Ben]:

$$\hat{\beta}_{ridge} = (X^TX + \lambda I)^{-1}X^TY$$

thus,

$$g(Y) = X(X^TX + \lambda I)^{-1}X^TY$$

and

$$\begin{aligned} df(g) =& (1/\sigma^2) trace(cov(Y, X(X^TX+\lambda I)^{-1}X^TY)) \\ &= (1/\sigma^2) trace(cov(Y, X(X^TX+\lambda I)^{-1}X^TY)) \\ &= (1/\sigma^2) trace(cov(Y, Y)X(X^TX+\lambda I)^{-1}X^T) \\ &= trace(X(X^TX+\lambda I)^{-1}X^T). \end{aligned}$$

ii. For lasso, I don't want you to derive the degrees of freedom. Instead, look over [?] 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.

Solution [Ben]:

By Stein's lemma, if Y is Normal and our fitting procedure g(Y) is almost differentiable then $\sum_{i=1}^{n} \frac{\partial f_i}{\partial y_i}$ is an unbiased estimate of the degrees of freedom of g(Y). The paper first justifies that the lasso fit is indeed almost differentiable using the geometry of complex polyhedrons- specifically the lasso fit is almost differentiable due to it being the residual of the projection onto a complex polyhedron. Then, the author states the lasso problem in terms of the active set of the lasso solution (the active set is the set of indeces of the variables whose beta coefficients are not set to zero for a particular lasso solution). Although the active set of a solution is not unique, since the lasso solution itself is not unique, the author shows that the column space of the active x variables is unique, and further, that the divergence of the lasso fit is a function of this column space, making the lasso degrees of freedom problem well defined. The divergence is shown to be the rank of the matrix of active x variables, and so the expectation of this rank over the domain of Y is the degrees of freedom for the lasso problem.

d. **Generalized information criterion (GIC).** The original proposed GIC was in [?] 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

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

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.

i. **Solution [Henry]:** These choices work when $n >> p$. However, when $n \leq p$, this doesn't work at all because the MLE $\hat{\sigma}^2_\alpha$ will be zero. This is because the orthogonal complement of $\mathbb{X}$ is empty. We won't be able to estimate σ^2 because we will have projected Y into a larger space. In this case, the criterion doesn't help us at all. We will always choose the model with all parameters included.

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 [?] (NOTE: I think Darren actually means this reference: [?])) for more information). Just use the true variance for $\hat{\sigma}^2$ right now, but know that this is still a very open, interesting area of research (see [?] for a review). Note that $\kappa_n = 2$ corresponds to AIC with Gaussian errors, but assuming that the variance is known. Using the simulation in `1_simulation.tar`, compare the prediction risk for

iia. **Solution (It should probably be noted here that it was Henry's job to do this...):** It appears to the author that involved modification is required to use this code to simulate prediction risk for Ridge regression, since it is built for Lasso.

iib. **Solution (This incomplete solution is also Henry's fault):** Results for CV, AIC, and BIC are shown in the accompanying manuscript, "manuscriptInfoCriteriaSimulation.pdf". If one wished to investigate CCV as well, the necessary code already exists in '`1resultsF.R`'. To get at it, one would need to re-run at least a portion of the simulation, but could then use the provided plotting code to make additional violin plots. Sadly, the author himself was unable to overcome errors associated with the parallelization aspects of the code, and so generated no such violin plots. He wishes all other lots of luck, and is happy to share what he's learned, however incomplete.