STA 712: HW 2

Due: Friday, September 15, 12:00pm (noon) on Canvas

Instructions: Submit your work as a single PDF. Your document should be created using LaTeX; see the course website for a homework template file and instructions on getting started with LaTeX and Overleaf.

1 Two views of AIC

AIC and the KL divergence

Let Y denote the response variable of interest in a generalized linear model, and let $f_0(y)$ denote the *true* probability function of the response (i.e., the true distribution that generated the data). For simplicity, assume here that Y is continuous, so $f_0(y)$ is a density.

We consider a family of distributions $\{f_\theta(y) : \boldsymbol{\theta} \in \Theta\}$, parameterized by a vector $\boldsymbol{\theta} \in \mathbb{R}^p$. Our goal, when building a model, is to choose the parameters $\boldsymbol{\theta}$ which make $f_\theta(y)$ as close to the *true* distribution $f_0(y)$ as possible.

One method of measuring how close $f_\theta(y)$ is to $f_0(y)$ is with the **Kullback-Leibler** (KL) divergence:

$$K(f_\theta, f_0) = \int \log\left(\frac{f_0(y)}{f_\theta(y)}\right) f_0(y) dy. \tag{1}$$

We fit our model via maximum likelihood estimation, producing the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ (which depends on the observed data). Then, $K(f_{\widehat{\theta}}, f_0)$ is one measure of how close our model is expected to fit a *new* set of data. And one way of choosing a model would be to choose the model which does best at predicting new data, i.e. which minimizes the KL divergence.

In the following questions, you will explore some basics of KL divergence, and show that choosing a model to minimize AIC is equivalent to minimizing an estimate of the KL divergence.

1. Show that $K(f_\theta, f_0) = 0$ if $f_\theta = f_0$.
2. Using the fact that $log(x) \leq x - 1$ for all x, show that $K(f_\theta, f_0) \geq 0$ if $f_\theta \neq f_0$.
3. Conclude that, if we *could* calculate KL divergence for our models, and the true model f_0 was among the models considered in the model selection procedure, then minimizing KL divergence would find the true model.
4. Explain why we can't actually calculate KL divergence when modeling real data.

Since we can't calculate KL divergence, we to approximate it. Let $\boldsymbol{\theta}^*$ be the value of $\boldsymbol{\theta}$ which minimizes $K(f_\theta, f_0)$. A Taylor expansion gives

$$K(f_{\widehat{\theta}}, f_0) \approx K(f_{\theta^*}, f_0) + \frac{1}{2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T \mathcal{I}(\boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*), \tag{2}$$

where $\mathcal{I}(\boldsymbol{\theta}^*)$ is the information matrix at $\boldsymbol{\theta}^*$.

1. We know from STA 711 that, if the model is correct and the sample size is sufficiently large, then $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T \mathcal{I}(\boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \approx \chi^2_p$. Conclude that under these assumptions,

$$\mathbb{E}[K(f_{\widehat{\theta}}, f_0)] \approx K(f_{\theta^*}, f_0) + p/2. \tag{3}$$

So far, our approximation to $K(f_{\widehat{\theta}}, f_0)$ is $K(f_{\theta^*}, f_0) + p/2$ (we can already see a penalty term that involves the number of parameters!). However, $K(f_{\theta^*}, f_0)$ still depends on f_0. To estimate $K(f_{\theta^*}, f_0)$, let $\boldsymbol{Y} = [Y_1, ..., Y_n]^T$ denote the observed data used to calculate $\widehat{\boldsymbol{\theta}}$, and consider the log-likelihood $\ell(\widehat{\boldsymbol{\theta}}) = -\sum_i \log f_{\widehat{\theta}}(Y_i)$.

1. Under the same assumptions as the previous question, $2(\ell(\widehat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}^*)) \approx \chi^2_p$. Use this to argue that

$$K(f_{\theta^*}, f_0) \approx \mathbb{E}[-\ell(\widehat{\boldsymbol{\theta}})] + p/2 + \int \log(f_0(y)) f_0(y) dy. \tag{4}$$

2. Combine the previous questions to argue that an estimate of $K(f_{\widehat{\theta}}, f_0)$ is

$$\widehat{K(f_{\widehat{\theta}}, f_0)} = -\ell(\widehat{\boldsymbol{\theta}}) + p + \int \log(f_0(y)) f_0(y) dy. \tag{5}$$

3. Conclude that minimizing $\widehat{K(f_{\widehat{\theta}}, f_0)}$ is equivalent to minimizing

$$AIC = -2\ell(\widehat{\boldsymbol{\theta}}) + 2p.$$

This derivation isn't super rigorous, but it captures the main steps and the important intuition: AIC is a reasonable metric for model selection because minimizing AIC is equivalent to minimizing an estimate of the KL divergence.

AIC and LOOCV

Another way of viewing AIC is as a computationally efficient approximation of leave-one-out cross-validation (LOOCV). Indeed, Stone (1977) showed that (under appropriate assumptions), *AIC and LOOCV are asymptotically equivalent.* Stone's paper is quite short (indeed, it is actually published in the journals *Notes, Queries, and Comments* rather than as a full-length research article), and this part of the assignment will guide you through reading the original paper.

> Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 44-47.

Begin by reading sections 1 – 3 of the paper, which set up the background and problem the author is trying to solve. (Because this is a short research note, rather than a full paper, the structure of the article is a bit different than usual).

1. In this paper, what is the LOOCV metric that we want to minimize? That is, what value is calculated for each held-out observation?
2. The goal of the article is to show that model selection by AIC and LOOCV are asymptotically equivalent. In particular, which two quantities (give the equation numbers) does the author plan to show are equivalent?

Now read section 4 (you may need to read it several times). For your first read-through, I suggest skimming over details like regularity conditions, and focusing on the general outline of the mathematical results. In particular:

1. What technique (which we have used many times in class!) does the author use to characterize the large-sample behavior of A?
2. What is the key assumption that makes A asymptotically equivalent to AIC?

Finally, read through section 4 again, keeping the outline and direction of the derivation in mind as you read. This time, pay more attention to the mathematical details and see if you can follow the steps. The key equations are (4.4), (4.5), and (4.6).

1. What asymptotic results are used to move from equation (4.4) to equation (4.5)?
2. How does the "key assumption" allow us to write equation (4.5) as equation (4.6)? In particular, what is required for $L_1 = -L_2$? (We have seen this equivalence before in STA 711!)

Reflecting

In this assignment, we have seen two justifications of AIC. First, using AIC is reasonable because minimizing AIC is the same as minimizing an estimate of the KL divergence. And minimizing an estimate of the KL divergence is reasonable because *if* the true model is contained in our search, and *if* we could minimize the actual KL divergence (not the estimate) then we would choose the true model. Second, minimizing AIC is reasonable because it is asymptotically equivalent to minimizing LOOCV error. As long as we think LOOCV error is a good method of assessing model performance, then AIC should be too.

Let's finish the assignment by comparing these two different perspectives of the AIC.

1. What are the similarities between the two perspectives of AIC (KL divergence and LOOCV)? Think about:

   - What KL divergence and LOOCV are trying to measure
   - Similar methods used in both perspectives

   Provide an intuitive explanation for why the LOOCV error in the Stone (1977) paper (the log likelihood calculated with each training observation held out in turn) would be similar to the KL divergence (which involves expected log-likelihoods...)