Stat 151a Lecture Notes 20

Chae Yeon Lee

April 25, 2023

1 Overview

Today:

- Finish Model Selection
- Introduction to Shrinkage Methods

Relevant Readings:

- Fox 22.2.1
- James 6.2

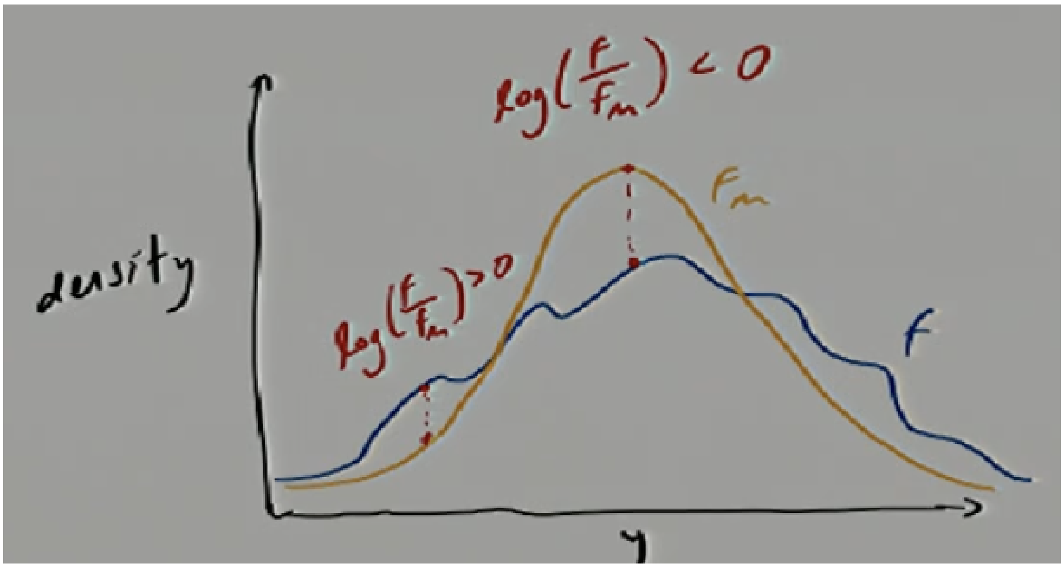
2 AIC, BIC

These are more classical model selection criteria that you will often see. **AIC** is based on a quantity called the **KL divergence**.

$$D_{KL}(f, f_m) = \int_{domain} log(\frac{f(y)}{f_m(y)}) f(y) dy$$

where f is the true density of (x, y) and f_m is the density under the model m.
Notes:

- D_{KL} is quite related to entropy
- It is a principled way of measuring distance from the true distribution, f.
- We think of f_m as an approximation to f for our model, m, with parameters, $\vec{\theta}$.



Stated as Fact:

1. $D_{KL}(f, f_m) \geq 0 \; \forall$ distribution f, f_m KL divergence will be a positive value for all distributions because if the model likelihood is high at some points, it must be lower at other points for the integral to integrate to 1.
2. $D_{KL}(f, f_m) = 0 \Leftrightarrow f(y) = f_m(y) \forall y \in Y$ The KL divergence is zero only if the distributions are perfectly equal.
3. $D_{KL}(f, f_m) \neq D_{KL}(f_m, f)$ usually.
4. KL divergence is a weighted (by $f(y)$) average of $log(\frac{f}{f_m})$.(i.e. $f(y)$ is large and $f_m(y)$ is small for some $y \in Y$, then this is a large positive addition to the divergence.)

Fact: $D_{KL}(f, f_m) = C - \mathbb{E}[log(f_m(y))]$ where c is the function of $f(y)$.
Proof:

$$D_{KL}(f, f_m) = \int log(\frac{f(y)}{f_m(y)} f(y) dy$$

$$= \int log(f(y)) f(y) dy - \int log(f_m(y)) f(y) dy$$

$$= C - \mathbb{E}[log(f_m(y))]$$

So, if we can estimate $\mathbb{E}[log(f_m(y))]$, then we can use that to compare models.
Rough Sketch of how to estimate the target quantity.

$$\mathbb{E}[log(f_m(y))] \simeq \frac{1}{n} \sum_{i=1}^{n} log(f_m(y_i))$$

(step a)

$$= \frac{1}{n} log(\prod_{i=1}^{n} f_m(y_i))$$

$$= \frac{1}{n} logL(\hat{\beta_m}|y_1, y_2, \cdots, y_n)$$

$$\simeq \frac{1}{n} logL(\hat{\beta_m}|y_1, \cdots, y_n)$$

(Step D) The MLE given $y_1, \cdots, y_n$ is the best approximation to $\vec{\beta_m}$.
Problem: We've used the same data to approximate the initial expectation as well as the model likelihood, which overstates the expected log likelihood of f_m over f.

$$\mathbb{E}[D_{KL}(f, f_m)] \approx C - logL(\hat{\beta}_m|\vec{y}) + (P_m + 1)$$

The expected value of KL divergence is approximated with log likelihood plus number of covariates in model L. Thus, the number of covariates increases the expected value of KL divergence. We need to **penalize** the likelihood by the number of covariates.
This derivate is omitted because it is quite technical.
Thus, argmin $\{D_{KL}(f, f_m)\} \approx$ argmin $\{-logL(\hat{\beta}_m|\vec{y}) + (P_m + 1)\}$ where $\{-logL(\hat{\beta}_m|\vec{y}) + (P_m + 1)\}$ is equivalent to $\frac{1}{2} AIC(m)$.
AIC = $-2logL(\hat{\beta_m}|\vec{y}) + 2(P_m + 1)$
BIC = $-2logL(\hat{\beta_m}|\vec{y}) + (log(n))(P_m + 1)$
The reason that we multiply by 2 is to make the AIC more comparable to the BIC criteria.
We don't derive this but the BIC derivation is from a Bayesian representation of model likelihood (rather than the KL divergence).
Note: For both AIC and BIC, a smaller value is a better model.

3 Trade offs and Selection in Practice

Keep in mind that for AIC, BIC, and MSE of cross validation. We want to pick a model that minimizes these criteria. We should be $\underline{careful to}$: treat categorical dummies jointly, and optionally, enforce: the principle of marginality.
Tradeoffs of using AIC and/or BIC versus Cross Validation
Pro:

- Only train each model once.
- Use all the data at once.

Con:

- Are based on a likelihood model and can be difficult to compare AIC and BIC over model types (nonlinear, non-Gaussian models).
- BIC is more parsimonious than AIC. BIC's $log(n)$ parameter vs 2 for AIC. For most of the reasonable sized data, BIC is more preferable.

In today's age, if we have K total features under consideration, and K is reasonably sized, then we can run a different model for subsets of the k features, and pick the model with the lowest AIC / BIC.