Lecture Notes 16
Model Selection

Not in the text except for a brief mention in 13.6.

1 Introduction

Sometimes we have a set of possible models and we want to choose the best model. Model selection methods help us choose a good model. Here are some examples.

Example 1 *Suppose you use a polynomial to model the regression function:*

$$m(x) = \mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p.$$

You will need to choose the order of polynomial p. We can think of this as a sequence of models $\mathcal{M}_1, \ldots, \mathcal{M}_p, \ldots$ indexed by p.

Example 2 *Suppose you have data $Y_1, \ldots, Y_n$ on age at death for n people. You want to model the distribution of Y. Some popular models are:*

1. *$\mathcal{M}_1$: the exponential distribution: $p(y;\theta) = \theta e^{-\theta y}$.*
2. *$\mathcal{M}_2$: the gamma distribution: $p(y;a,b) = (b^a/\Gamma(a))y^{a-1}e^{-by}$.*
3. *$\mathcal{M}_3$: the log-normal distribution: we take $\log Y \sim N(\mu, \sigma^2)$.*

Example 3 *Suppose you have time series data $Y_1, Y_2, \ldots$. A common model is the AR (autoregressive model):*

$$Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + \cdots + a_k Y_{t-k} + \epsilon_t$$

where $\epsilon_t \sim N(0, \sigma^2)$. The number k is called the order of the model. We need to choose k.

Example 4 *In a linear regression model, you need to choose which variables to include in the regression. This is called variable selection. This problem is discussed at length in 36-707 and 10/36-702.*

Example 5 *A commonly used model is the mixture-of-Gaussians:*

$$p(y) = \sum_{j=1}^{k} \pi_j \, \phi(y;\ \mu_j, \Sigma_j)$$

where $\pi_j \geq 0$, $\sum_j \pi_j = 1$ and ϕ denotes a Multivariate Normal density. The model selection problem here is to choose the number of components k.

The most common model selections methods are:

1. AIC (and related methods like C_p).
2. Cross-validation.
3. BIC (and related methods like MDL, Bayesian model selection).

We need to distinguish between 2 goals:

1. Find the model that gives the best prediction (without assuming that any of the models are correct).
2. Assume one of the models is the true model and find the "true" model.

Generally speaking, AIC and cross-validation are used for goal 1 while BIC is used for goal 2.

2 AIC

Suppose we have models $\mathcal{M}_1, \ldots, \mathcal{M}_k$ where each model is a set of densities:

$$\mathcal{M}_j = \left\{ p(y;\theta_j) : \ \theta_j \in \Theta_j \right\}.$$

We have data $Y_1, \ldots, Y_n$ drawn from some density f. **We do not assume that f is in any of the models.**

Let $\widehat{\theta}_j$ be the mle from model j. An estimate of p, based on model j is $\widehat{p}_j(y) = p(y;\widehat{\theta}_j)$. The quality of $\widehat{p}_j(y)$ as an estimate of f can be measured by the Kullback-Leibler distance:

$$\begin{aligned} K(p,\widehat{p}_j) &= \int p(y) \log\left(\frac{p(y)}{\widehat{p}_j(y)}\right) dy \\ &= \int p(y)\log p(y) dy - \int p(y)\log \widehat{p}_j(y) dy. \end{aligned}$$

The first term does not depend on j. So minimizing $K(p,\widehat{p}_j)$ over j is the same as maximizing

$$K_j = \int p(y)\log p(y;\widehat{\theta}_j) dy.$$

We need to estimate K_j. Intuitively, you might think that a good estimate of K_j is

$$\overline{K}_j = \frac{1}{n}\sum_{i=1}^{n} \log p(Y_i;\widehat{\theta}_j) = \frac{\ell_j(\widehat{\theta}_j)}{n}$$

where $\ell_j(\theta_j)$ is the log-likelihood funcion for model j. However, this estimate is very biased because the data are being used twice: first to get the mle and second to estimate the

integral. Akaike showed that the bias is approximately d_j/n where $d_j = \text{dimension}(\Theta_j)$. Therefore we use

$$\widehat{K}_j = \frac{\ell_j(\widehat{\theta}_j)}{n} - \frac{d_j}{n} = \overline{K}_j - \frac{d_j}{n}.$$

Now, define

$$\text{AIC}(j) = 2n\widehat{K}_j = \ell_j(\widehat{\theta}_j) - 2d_j.$$

Notice that maximizing $\widehat{K}_j$ is the same as maximizing $\text{AIC}(j)$ over j. Why do we multiply by $2n$? Just for historical reasons. We can multiply by any constant; it won't change which model we pick. In fact, different texts use different versions of AIC.

AIC stands for "Akaike Information Criterion." Akaike was a famous Japanese statistician who died recently (August 2009).

3 Theoretical Derivation of AIC (Optional)

Let us now look closer to see where the formulas come from. Recall that

$$K_j = \int p(y) \log p(y; \widehat{\theta}_j) dy.$$

For simplicity, let us focus on one model and drop the subscript j. We want to estimate

$$K = \int p(y) \log p(y; \widehat{\theta}) dy.$$

Our goal is to show that

$$\overline{K} - \frac{d}{n} \approx K$$

where

$$\overline{K} = \frac{1}{n} \sum_{i=1}^{n} \log p(Y_i; \widehat{\theta})$$

and d is the dimension of θ.

Some Notation and Background. Let θ_0 minimize $K(f, p(\cdot; \theta))$. So $p(y; \theta_0)$ is the closest density in the model to the true density. Let $\ell(y, \theta) = \log p(y; \theta)$ and

score = gradient of the log-likelihood w.r.t. the parameter vector

$$s(y, \theta) = \frac{\partial \log p(y; \theta)}{\partial \theta}$$

be the score and let $H(y, \theta)$ be the matrix of second derivatives.

Let $Z_n = \sqrt{n}(\widehat{\theta} - \theta_0)$ and recall that

unbiased?

$$Z_n \rightsquigarrow N(0, J^{-1} V J^{-1})$$

where $J = -\mathbb{E}[H(Y, \theta_0)]$ and

$$V = \text{Var}(s(Y, \theta_0)).$$

In class we proved that $V = J^{-1}$. But that proof assumed the model was correct. We are not assuming that. Let

$$S_n = \frac{1}{n}\sum_{i=1}^{n} s(Y_i, \theta_0).$$

By the CLT,

$$\sqrt{n}S_n \rightsquigarrow N(0, V)$$

Hence, in distribution

$$JZ_n \approx \sqrt{n}S_n. \tag{1}$$

Here we used the fact that $\text{Var}(AX) = A(\text{VarX})A^T$. Thus

$$\text{Var}(JZ_n) = J(J^{-1}VJ^{-1})J^T = V.$$

We will need one other fact. Let ϵ be a random vector with mean μ and covariance Σ. Let

$$Q = \epsilon^T A \epsilon.$$

(Q is a called a quadratic form.) Then

$$\mathbb{E}(Q) = \text{trace}(A\Sigma) + \mu^T A \mu.$$

The details. By using a Taylor series

$$\begin{aligned} K &\approx \int p(y)\left(\log p(y;\theta_0) + (\widehat{\theta} - \theta_0)^T s(y, \theta_0) + \frac{1}{2}(\widehat{\theta} - \theta_0)^T H(y, \theta_0)(\widehat{\theta} - \theta_0)\right) dy \\ &= K_0 - \frac{1}{2n} Z_n^T J Z_n \end{aligned}$$

where

$$K_0 = \int p(y) \log p(y; \theta_0) dy,$$

The second term dropped out because, like the score function, it has mean 0. Again we do a Taylor series to get

$$\begin{aligned} \overline{K} &\approx \frac{1}{n}\sum_{i=1}^{n}\left(\ell(Y_i, \theta_0) + (\widehat{\theta} - \theta_0)^T s(Y_i, \theta_0) + \frac{1}{2}(\widehat{\theta} - \theta_0)^T H(Y_i, \theta_0)(\widehat{\theta} - \theta_0)\right) \\ &= K_0 + A_n + (\widehat{\theta} - \theta_0)^T S_n - \frac{1}{2n} Z_n^T J_n Z_n \\ &\approx K_0 + A_n + \frac{Z_n^T S_n}{\sqrt{n}} - \frac{1}{2n} Z_n^T J Z_n \end{aligned}$$

where

$$J_n = -\frac{1}{n}\sum_{i=1}^{n} H(Y_i, \theta_0) - \xrightarrow{P} J,$$

and

$$A_n = \frac{1}{n}\sum_{i=1}^{n}(\ell(Y_i, \theta_0) - K_0).$$

Hence,

$$\overline{K} - K \approx A_n + \frac{\sqrt{n} Z_n^T S_n}{n} \approx A_n + \frac{Z_n^T J Z_n}{n}$$

where we used (1). We conclude that

$$\mathbb{E}(\overline{K} - K) \approx \mathbb{E}(A_n) + \mathbb{E}\left(\frac{Z_n^T J Z_n}{n}\right) = 0 + \frac{\text{trace}(J\ J^{-1} V J^{-1})}{n} = \frac{\text{trace}(J^{-1}V)}{n}.$$

Hence,

$$K \approx \overline{K} - \frac{\text{trace}(J^{-1}V)}{n}.$$

If the model is correct, then $J^{-1} = V$ so that $\text{trace}(J^{-1}V) = \text{trace}(I) = d$. Thus we would use

$$K \approx \overline{K} - \frac{d}{n}.$$

You can see that there are a lot of approximations and assumptions being used. So AIC is a very crude tool. Cross-validation is much more reliable.

4 Cross-Validation

There are various flavors of cross-validation. In general, the data are split into a training set and a test set. The models are fit on the training set and are used to predict the test set. Usually, many such splits are used and the result are averaged over splits. However, to keep things simple, we will use a single split.

Suppose again that we have models $\mathcal{M}_1, \ldots, \mathcal{M}_k$. Assume there are $2n$ data points. Split the data randomly into two halves that we will denote $D = (Y_1, \ldots, Y_n)$ and $T = (Y_1^*, \ldots, Y_n^*)$. Use D to find the mle's $\widehat{\theta}_j$. Then define

$$\widehat{K}_j = \frac{1}{n}\sum_{i=1}^{n} \log p(Y_i^*; \widehat{\theta}_j).$$

Note that $\mathbb{E}(\widehat{K}_j) = K_j$; there is no bias because $\widehat{\theta}_j$ is independent of Y_j^*. We will assume that $|\log p(y;\theta)| \leq B < \infty$. By Hoeffding's inequality,

$$\mathbb{P}(\max_j |\widehat{K}_j - K_j| > \epsilon) \leq 2ke^{-2n\epsilon^2/(2B^2)}.$$

Let

$$\epsilon_n = \sqrt{\frac{2B^2 \log(2k/\alpha)}{n}}.$$

Then

$$\mathbb{P}(\max_j |\widehat{K}_j - K_j| > \epsilon_n) \leq \alpha.$$

If we choose $\widehat{j} = \text{argmax}_j \widehat{K}_j$, then, with probability at least $1 - \alpha$,

$$K_{\widehat{j}} \geq \max_j K_j - 2\sqrt{\frac{2B^2 \log(2k/\alpha)}{n}} = \max_j K_j - O\left(\frac{\log k}{n}\right).$$

So with high probability, you choose close to the best model. This argument can be improved and also applies to regression, classification etc. Of course, with regression, the loss function is $\mathbb{E}(Y - m(X))^2$ and the cross-validation score is then

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i^* - m(X_i^*))^2.$$

For classification we use

$$\frac{1}{n}\sum_{i=1}^{n} I(Y_i^* \neq h(X_i^*)).$$

We have made essentially no assumptions or approximations. (The bounded on $\log f$ can be relaxed.) The beauty of cross-validation is its simpicity and generality. It can be shown that AIC and cross-validation have very similar behavior. But, cross-validation works under weaker conditions.

5 BIC

BIC stands for *Bayesian Information Criterion*. It is also known as *the Schwarz Criterion* after Gideon Schwarz. It is virtually identical to the MDL (minimum description length) criterion.

We choose j to maximize

$$\text{BIC}_j = \ell_j(\widehat{\theta}_j) - \frac{d_j}{2}\log n.$$

This is the same as AIC but the penalty is harsher. Thus, BIC tends to choose simpler models. Here is the derivation.

We put a prior $\pi_j(\theta_j)$ on the parameter θ_j. We also put a prior probability p_j that model $\mathcal{M}_j$ is the true model. By Bayes theorem

$$P(\mathcal{M}_j|Y_1, \ldots, Y_n) \propto p(Y_1, \ldots, Y_n|\mathcal{M}_j)p_j.$$

Furthermore,

$$p(Y_1, \ldots, Y_n|\mathcal{M}_j) = \int p(Y_1, \ldots, Y_n|\mathcal{M}_j, \theta_j)\pi_j(\theta_j)d\theta_j = \int L(\theta_j)\pi_j(\theta_j)d\theta_j.$$

We know choose j to maximize $P(\mathcal{M}_j|Y_1, \ldots, Y_n)$. Equivalently, we choose j to maximize

$$\log \int L(\theta_j)\pi_j(\theta_j)d\theta_j + \log p_j.$$

Some Taylor series approximations show that

$$\log \int L(\theta_j)\pi_j(\theta_j)d\theta_j + \log p_j \approx \ell_j(\widehat{\theta}_j) - \frac{d_j}{2}\log n = \text{BIC}_j.$$

What happened to the prior? It can be shown that the terms involving the prior are lower order than the term that appear in formula for BIC_j so they have been dropped.

BIC behaves quite differently than AIC or cross-validation. It is also based on different assumptions. BIC assumes that one of the models is true and that you are trying to find the model most likely to be true in the Bayesian sense. AIC and cross-validation are trying to find the model that predict the best.

6 Model Averaging

Bayesian Approach. Suppose we want to predict a new observation Y. Let $D = \{Y_1, \ldots, Y_n\}$ be the observed data. Then

$$p(y|D) = \sum_j p(y|D, \mathcal{M}_j)\mathbb{P}(\mathcal{M}_j|D)$$

where

$$\mathbb{P}(\mathcal{M}_j|D) = \frac{\int L(\theta_j)\pi_j(\theta_j)d\theta_j}{\sum_s \int L(\theta_s)\pi_s(\theta_s)d\theta_s} \approx \frac{e^{\text{BIC}_j}}{\sum_s e^{\text{BIC}_s}}.$$

Frequentist Approach. There is a large and growing literaure on frequentist model averaging. We don't have time to study it here.

7 Simple Normal Example

Let

$$Y_1, \ldots, Y_n \sim N(\mu, 1).$$

We want to compare two models:

$$M_0 : \ N(0, 1), \quad \text{and} \quad M_1 : \ N(\mu, 1).$$

Hypothesis Testing. We test

$$H_0 : \mu = 0 \quad \text{versus} \quad \mu \neq 0.$$

The test statistic is

$$Z = \frac{\overline{Y} - 0}{\sqrt{\mathsf{Var}(\overline{Y})}} = \sqrt{n}\,\overline{Y}.$$

We reject H_0 if $|Z| > z_\alpha/2$. For $\alpha = 0.05$, we reject H_0 if $|Z| > 2$, i.e., if

$$|\overline{Y}| > \frac{2}{\sqrt{n}}.$$

AIC. The likelihood is proportional to

$$\mathcal{L}(\mu) = \prod_{i=1}^{n} e^{-(Y_i-\mu)^2/2} = e^{-n(\overline{Y}-\mu)^2/2} e^{-nS^2/2}$$

where $S^2 = \sum_i (Y_i - \overline{Y})^2$. Hence,

$$\ell(\mu) = -\frac{n(\overline{Y}-\mu)^2}{2} - \frac{nS^2}{2}.$$

Recall that $AIC = \ell_S - |S|$. The AIC scores are

$$AIC_0 = \ell(0) - 0 = -\frac{n\overline{Y}^2}{2} - \frac{nS^2}{2}$$

and

$$AIC_1 = \ell(\widehat{\mu}) - 1 = -\frac{nS^2}{2} - 1$$

since $\widehat{\mu} = \overline{Y}$. We choose model 1 if

$$AIC_1 > AIC_0$$

that is, if

$$-\frac{nS^2}{2} - 1 > -\frac{n\overline{Y}^2}{2} - \frac{nS^2}{2}$$

or

$$|\overline{Y}| > \frac{\sqrt{2}}{\sqrt{n}}.$$

Similar to but not the same as the hypothesis test.

BIC. The BIC scores are

$$BIC_0 = \ell(0) - \frac{0}{2}\log n = -\frac{n\overline{Y}^2}{2} - \frac{nS^2}{2}$$

and

$$BIC_1 = \ell(\widehat{\mu}) - \frac{1}{2}\log n = -\frac{nS^2}{2} - \frac{1}{2}\log n.$$

We choose model 1 if

$$BIC_1 > BIC_0$$

that is, if

$$|\overline{Y}| > \sqrt{\frac{\log n}{n}}.$$

Hypothesis testing	controls type I errors
AIC/CV/C_p	finds the most predictive model
BIC	finds the true model (with high probability)