Lecture Notes 25
36-705

Today we will discuss the problem of *model selection*. Let's start with some examples.

Example 1: A convenient, flexible parametric family is the mixture of Gaussians:

$$p_\theta(x) = \sum_{i=1}^{k} \pi_i N(\mu_i, \Sigma_i)$$

where $\sum_i \pi_i = 1$. The parameters are $\theta = (\pi_1, \ldots, \pi_k, \mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k)$. We also need to choose the number of mixture components k and this is a model selection problem where we have a sequence of models $\mathcal{M}_1, \ldots, \mathcal{M}_k$ indexed by the number of components.

Example 2: Polynomial order in regression. 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 3: AR model. We have always assumes that the data are iid. An example of non iid data is a time series where we expect the data to be correlated over time. Consider a time series $Y_1, Y_2, \ldots$. 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 survival analysis we want to model lifetime Y which is a non-negative random variable. Two common models are the Weibull distribution $p_\theta(x) = (k/\lambda)(x/\lambda)^{k-1} e^{-(x/\lambda)^k}$ and the lognormal, where $\log X \sim N(\mu, \sigma^2)$. This defines two different models $\mathcal{M}_1$ and $\mathcal{M}_2$.

Notice that models are often nested with increasing complexity. Choosing a large k can lead to overfitting. Just picking the model with the highest likelihood will lead to always picking the biggest model. There are two different goals in model selection:

1. Find the model that gives the best prediction (without assuming that any of the models are correct). This is equivalent to: find the model whose estimated distribution is closest to the true distribution.

2. Assume one of the models is the true model (the smallest model containing the true density) and find the true model.

We'll talk about the following methods:

Cross-validation
AIC
BIC
Bayes.

Perhaps the only slightly counter-intuitive fact you need to remember is that when there is a true model methods like cross-validation can fail to find it.

The basic take-aways are:

1. If your goal is prediction, you have a reasonable sample-size and you have a reasonable computation budget use cross-validation.

2. If your goal is prediction, but you either have too small a sample or you have a very low computational budget, you should consider using AIC.

3. If your goal is selecting the true model you should use BIC.

1 The Methods

CV: Cross-validation has different versions but the procedure is probably roughly familiar to you. We train our models on a subset of the data and then evaluate and choose between the models on the rest of the data. We then potentially re-shuffle the data, repeat, and combine the results in some way. We will simplify this and just suppose throughout this lecture that we do a train-test split.

AIC: AIC (Akaike Information Criterion) is a model selection rule that does not use any sample-splitting. Informally, it is best understood as an asymptotic approximation to CV. Formally, Stone showed in a classic paper that under assumptions AIC and CV are asymptotically equivalent when using the MLE for each model.

BIC. BIC (Bayesian Information Criterion) can be thought of as an asymptotic approximation of a Bayesian approach.

2 Cross Validation

Denote the sample size by $2n$. Split the data into two groups $\mathcal{D}_1$ and $\mathcal{D}_2$. (In practice we often use many groups.) Using $\mathcal{D}_1$ we find the mle $\widehat{\theta}_j$ for model $\mathcal{M}_j$. Let $p_j(x) = p(x; \widehat{\theta}_j)$ be te estimated density from model $\mathcal{M}_j$. The idea is to choose j which minimizes $K(p, p_j)$

where p is the true density and

$$K(p, p_j) = \int p(x) \log \left(\frac{p(x)}{p_j(x)} \right) dx$$

is the Kullback-Leibler distance. Notice that we do not necessarily assume that the true density p is in any of the models.

Notice that minimizing $K(p, p_j)$ is the same as maximizing

$$R_j = \int p(x) \log p_j(x) dx.$$

We can estimate R_j from the second sample $\mathcal{D}_2$ by

$$\widehat{R}_j = \frac{1}{n} \sum_{i \in \mathcal{D}_2} \log p(X_i; \widehat{\theta}_j).$$

The score $\widehat{R}_j$ can also be thought of as a measure of how well we can predict future observations. If we did not split the data, $\widehat{R}_j$ would be biased: larger models will always lead to a larger score.

We can now use the LLN to argue that if the test-set size goes to ∞ then our risk estimates converge to their expectations, and then we will find the model/estimate with the lowest KL to the true model. Let's make this more precise. Assume that $|\log p_\theta(X)| \leq B$ for every θ and X that we care about (this can be relaxed using more complex techniques). From Hoeffding's inequality and the union bound:

$$\mathbb{P}(\max |R_i - \mathbb{E}(R_i)| \geq \epsilon) \leq 2M \exp(-2n\epsilon^2/(4B^2)).$$

Let

$$\epsilon_n = \sqrt{\frac{4B^2 \log(2M/\alpha)}{n}}.$$

Then

$$\mathbb{P}(\max_i |R_i - \mathbb{E}(R_i)| \geq \epsilon_n) \leq \alpha.$$

Let $\widehat{i} = \arg\min_i R_i$ be the selected model and let $i^* = \arg\min_i \mathbb{E}(R_i)$ be the best model. Then, with probability at least $1 - \alpha$:

$$\mathbb{E}(R_{\widehat{i}}) \leq R_{\widehat{i}} + \epsilon_n \leq R_{i^*} + \epsilon_n \leq \mathbb{E}(R_{i^*}) + 2\epsilon_n.$$

So the model we select will be sub-optimal by at most $2\epsilon_n$. In regression, we would use exactly the same reasoning, but just replace the risk with the squared loss. Reasoning about

K-fold cross-validation turns out to be much more challenging, because the data re-use breaks independence assumptions.

The analysis above should remind you of the analysis we did before of Empirical Risk Minimization. The goals are slightly different, as is the final guarantee. It is worth thinking about what exactly the data splitting buys you. In particular, we do not require uniform convergence of the empirical to the true risk over all the model classes $\mathcal{M}_1, \ldots, \mathcal{M}_M$, rather we only require a good estimate of the risk for the *fixed* models indexed by $\widehat{\theta}_1, \ldots, \widehat{\theta}_M$.

3 AIC

Suppose we don't want to use data-splitting because of lack of data or lack of computational resources. We could try to estimate R_j by

$$\widehat{R}_j = \frac{1}{n}\sum_i \log p(X_i; \widehat{\theta}_j) = (1/n)\ell_j(\widehat{\theta}_j)$$

but, as we discussed above, this is very biased. And the more parameters there are, ther more biased this will be. Akaike proved that the bias is approximately equal to d_j/n where d_j is the dimension of the model $\mathcal{M}_j$. So a bias-adjusted estimator is

$$\widehat{R}_j = (1/n)[\ell_j(\widehat{\theta}_j) - d_j].$$

For a variety of historical reasons, we will multiple by $2n$ which does not affect which model is the maximizer. This leads to the criterion

$$\mathrm{AIC}_j = 2\ell_j(\widehat{\theta}_j) - 2d_j.$$

We choose j to maximize AIC_j. It is best to think of this as an approximation to CV.

4 BIC

. The BIC crierion is

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

This is basically AIC but with a harsher penalty. This approach was proposed by Gideon Schwartz based on the following Bayesian argument. We place prior probabilities $\nu_1, \ldots, \nu_k$ for each model. Then we put priors $p_j(\theta_j)$ for the parameters in model $\mathcal{M}_j$. By Bayes' theorem

$$p(\mathcal{M}_j|X_1, \ldots, X_n) = \frac{p(X_1, \ldots, X_n|\mathcal{M}_j)\nu_j}{\sum_s p(X_1, \ldots, X_n|\mathcal{M}_s)\nu_s} = \frac{\nu_j \int \mathcal{L}_j(\theta_j)d\theta_j}{\sum_s \nu_s \int \mathcal{L}_s(\theta_s)d\theta_s}.$$

Schwartz showed that

$$\log p(\mathcal{M}_j|X_1,\ldots,X_n) \approx \text{BIC}_j.$$

Suppose that the true density is in one of the models. Let $\mathcal{M}_j$ be the smallest model that contains p. Then it can be shown that $\widehat{j} \xrightarrow{P} j$ where $\widehat{j}$ is the model chosen by BIC. We'll return to this point later.

5 Hypothesis Testing

In some cases we can frame model selection as a hypothesis testing problem. For example, suppose that $X_1,\ldots,X_n \sim N(\theta,1)$ and that are two models are

$$\mathcal{M}_0 = \{\theta : \theta = 0\},$$
$$\mathcal{M}_1 = \{\theta : \theta \in \mathbb{R}\}.$$

We can select a model by testing $H_0 : \theta = 0$ versus $H_0 : \theta = 1$. Notice that, if $\theta = 0$, we will choose the wrong model with probability α. But that's ok since CV and AIC also can choose the wrong model. But notice that hypothesis testing doesn't provide any guarantee about the selected model in terms of KL distance, for example.

6 Choosing the True Model?

Suppose that the true density is in one of the models. Let $\mathcal{M}_j$ be the smallest model that contains p. We have seen that, asymptotically, BIC chooses this model. Noe we will show that CV and AIC don't do this. But thus is not a problem since that's not their goal. We'll focus on the example above where the data re Normal and we are choosing between $\mathcal{M}_0$ and $\mathcal{M}_1$.

Let us denote the mean of the training samples as $\widehat{\mu}_{\text{tr}}$ and the mean of the testing samples as $\widehat{\mu}_{\text{te}}$. Then

$$\widehat{\theta}_0 = 0, \qquad \widehat{\theta}_1 = \widehat{\mu}_{\text{tr}}.$$

Focus on the case where $\theta = 0$. The difference between the cross-validation loss for the two models is (proportional to):

$$\frac{1}{n}\sum_{i=1}^n X_i^2 - \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{\mu}_{\text{tr}})^2 = -\widehat{\mu}_{\text{tr}}^2 + 2\widehat{\mu}_{\text{tr}}\widehat{\mu}_{\text{te}},$$

and we select the wrong model if this quantity is greater than 0. We can re-write this as: we select the wrong model if

$$(\sqrt{n}\widehat{\mu}_{\text{tr}})^2 - 2(\sqrt{n}\widehat{\mu}_{\text{tr}})(\sqrt{n})\widehat{\mu}_{\text{te}}) < 0.$$

Now we observe that, $\sqrt{n_{\text{tr}}}\widehat{\mu}_{\text{tr}}$ and $\sqrt{n_{\text{te}}}\widehat{\mu}_{\text{te}}$ are each independent $N(0,1)$ variables, so the probability of choosing the wrong model is

$$P(Z_1^2 - 2Z_1 Z_2 < 0)$$

where $Z_1, Z_2 \sim N(0,1)$. This is about 0.35 (no matter how large n is).

Again, this is not really a problem because choosing the true model is not the goal. Also, even if we choose the wrong model, $\widehat{\theta}_1$ will be close to 0 so $p(x;\widehat{\theta}_0)$ will be close to $N(0,1)$.

Now consider AIC. We select $\mathcal{M}_1$ if

$$\frac{1}{n}\sum_{i=1}^n X_i^2 \geq \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{\mu})^2 + \frac{1}{n}.$$

Re-arranging this we see that we would select the wrong model (i.e. Model 1) if,

$$\widehat{\mu}^2 \geq \frac{2}{n}.$$

This intuitively makes sense: if the mean is small in absolute value we select Model 0 and otherwise we select Model 1. Now $n\widehat{\mu}^2 \sim \chi_1^2$ so the probability of choosing $\mathcal{M}_1$ is $P(\chi_1^2 > 2) = 0.16$.

Now consder BIC. You can go through exactly the same calculation as above and see that BIC would select the wrong model if,

$$\widehat{\mu}^2 \geq \frac{\log n}{n},$$

which has probability $P(\chi_1^2 > \log n) \to 0$. If $\mu \neq 0$, you can check that the probability of choosing Model 1 goes to 1. Thus BIC is model seletion consistent. Thus is true of BIC more generally.

Finally, let us return to hypothesis testing. We would reject the null if

$$n\widehat{\mu}^2 \geq \chi_{1,\alpha}^2,$$

and this controls the Type I error (i.e. the error of incorrectly selecting the more complex model) at α. More generally, we could imagine testing between pairs of models using the LRT (using Wilks result for the asymptotic distributions). This procedure is very similar to AIC but inflates the penalty just enough to ensure that we have some specified error control.

7 What to do?

The derivations of AIC and BIC depend on many technical assumptions about the model. CV does not depend on any model assumptions. For this reason, CV is the better choice if it is feasible.

A difficult problem that we have not considered is how to account for model selection when doing inference. This is a complicated topic. The simplest think, if we have lots of data, is to keep some hold out data. After model selection, we use the hold out data which is not affected by the model selection process.

Another issue that we have not considered is interpretability. Getting good predictions is not the only goal. We might be willing to sacrifice a bit of prediction accuracy to have a more interpretable model. This is an area of active research.