36-705: Intermediate Statistics **Fall 2019**

Lecture 34: December 2

Lecturer: Siva Balakrishnan

Today we will discuss model selection. First, so that you have some canonical examples in mind:

Example 1: Suppose we are fitting a mixture of Gaussians, i.e. you observe data from some density and you want to estimate this density by a model of the form:

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

We then 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: Model order in AR. 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.

Notice that models are often nested, i.e. for instance you can perfectly model a mixture of $k-1$ Gaussians by a mixture of k Gaussians. Though this is not always the case, nested models should make the "over-fitting" problem clear, i.e. just picking the model with best "fit" would typically lead you to favor the most complex model.

There are two possible goals in model selection:

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.

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

The basic take-aways from today's lecture:

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.

34.1 The procedures

CV: Cross-validation has different versions but the procedure should be 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.

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

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

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

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

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

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

where $\ell_j(\theta_j)$ is the log-likelihood function 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

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

Notice that maximizing $\widehat{K}_j$ is the same as maximizing 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.

BIC: BIC (Bayesian Information Criterion) is similar to AIC except we use the criterion:

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

The main thing to observe is that BIC uses a harsher penalty for model complexity. Roughly in order to prefer a model with dimension $d+1$ to one of dimension d we need the log-likelihood to be much better for BIC while we only need it to be slightly better for AIC. In practice, BIC will tend to select much simpler/sparser/smaller models than AIC.

34.2 Prediction consistency of CV

Lets try to understand cross-validation in a simple scenario. We will not analyze AIC but you should note that in practice it is often quite similar to CV (and this can be formally shown under assumptions). We will do this in the context of point estimation, but one could use *exactly* the same argument for bandwidth selection.

Say we have models $\mathcal{M}_1, \ldots, \mathcal{M}_M$. These are different models that we think might be reasonable fits to the data. Now, we observe our data $(X_1, \ldots, X_{2n})$ and randomly split it into train and test sets of size n each. We really should refer to the test set as a validation set but we will ignore this for today.

On the train set, we fit our models (say using the MLE), and compute point estimates $\widehat{\theta}_1, \ldots, \widehat{\theta}_M$. Now, suppose that we want to select the model/estimate that fits the data well. We will use the negative log-likelihood as our measure, i.e., we want an estimate that has low negative log-likelihood. This is the same as using the KL divergence as our loss function.

We can use the test set to estimate the negative log-likelihood:

$$R_i = \frac{-1}{n} \sum_{i=1}^{n} \log f_{\widehat{\theta}_i}(X_{n+i}).$$

Note that:

$$\mathbb{E}(R_i) = -\mathbb{E}_{f_{\theta^*}} \log f_{\widehat{\theta}_i}(X) = KL(f_{\theta^*} || f_{\widehat{\theta}_i}) - \mathbb{E}_{f_{\theta^*}} \log f_{\theta^*}(X),$$

so we are estimating the KL divergence upto some term that does not depend on $\widehat{\theta}_i$. So minimizing $\mathbb{E}(R_i)$ is equivalent to minimizing the KL divergence.

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.

Suppose however we wanted to be more precise, and try to understand the role of the test set size and the number of models M?

We could use Hoeffding's inequality. This will need an assumption that $|\log f_\theta(X)| \leq B$ for every θ and X that we care about (this can be relaxed using more complex techniques). Now, notice that the following is an important but straightforward consequence of Hoeffding's inequality:

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

This is true since for each i we know that

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

so we can obtain the desired inequality via a union bound (if the max exceeds ϵ at least one of the terms must exceed ϵ).

Define,

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

then we know that

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

Suppose we select the model $\widehat{i} = \arg\min_i R_i$, and let $i^* = \arg\min_i \mathbb{E}(R_i)$, then we have that 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 $\widehat{\theta}_1, \ldots, \widehat{\theta}_M$.

34.3 Model selection

Now, we switch gears and consider the case when there is in fact a true model, and our goal is to select it (say with high-probability as the sample-size grows).

34.4 A basic test problem

Let us consider the simplest model selection problem. We observe samples

$$X_1, \ldots, X_n \sim N(\theta, 1),$$

and our two models are:

$$\mathcal{M}_1 = \{\theta : \theta = 0\},$$
$$\mathcal{M}_2 = \{\theta : \theta \in \mathbb{R}\}.$$

Our goal is to select one of these models (note that they are nested). Let us assume that $\theta = 0$ so the first model is indeed correct. We use the ℓ_2 loss. For an estimate $\widehat{\theta}$ the true prediction error is $\mathbb{E}(X - \widehat{\theta})^2 = \widehat{\theta}^2 + 1$.

Train error "overfits": The first thing to note is that using the train error results in us always selecting the wrong model. For the first model we use the estimate $\widehat{\theta}_1 = 0$ and for the second model we use the estimate that is $\widehat{\theta}_2 = \frac{1}{n}\sum_{i=1}^n X_i$.

Notice that with probability 1,

$$\frac{1}{n}\sum_{i=1}^n (X_i - \widehat{\theta}_2)^2 < \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{\theta}_1)^2,$$

so the train error would always select the more complex model (which in this case is also the wrong model).

34.5 Cross validation may not select the correct model

The next perhaps more surprising fact is that *no matter how large the sample size is*, there is a non-trivial chance that cross-validation does not select the correct model. Colloquially people say that cross-validation overfits or that cross-validation is not model-selection consistent.

We will again, analyze a simple train-test split version. We assume that the train and test sample sizes are each $n_{\text{tr}} = n_{\text{te}} = n/2$, and discuss this choice at the end.

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 notice that $\widehat{\theta}_2 = \widehat{\mu}_{\text{tr}}$, and further the difference between the test/cross-validation loss for the two models is simply:

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

so we select the wrong model if this quantity is greater than 0. Noting that we chose $n_{\text{tr}} = n_{\text{te}} = n/2$, we could re-write this as: we select the wrong model if

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

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 this happens with a reasonable probability (whenever the train and test averages share the same sign and the test average is at least half the train average). So more than 25% of the time CV will select the wrong model, no matter how large n is.

Observe that one could fix this problem by making the test set much larger than the train set, but you then sacrifice model fit (i.e. your estimate of θ will be quite bad) and as a result you will not predict as well. BIC is a much better way to fix cross-validation.

34.6 AIC also often selects the wrong model

Now, let us consider what AIC would do in our test problem. The log-likelihood in this case is simply half the ℓ_2 loss and so we select Model 2 if:

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

Re-arranging this we see that we would select the wrong model (i.e. Model 2) if,

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

This should intuitively make sense (if the mean is small in absolute value we select Model 1 and otherwise we select Model 2). The key point once again is that $n\widehat{\mu}^2$ is the square of

a standard normal variable, and is larger than 2 with some fixed probability (roughly 0.16). So once again no matter how large the sample size is AIC will select the wrong model with a fixed probability, i.e. it is not model selection consistent.

Roughly, the problem is that there is a penalty for selecting a more complex model but it is not "strong enough". BIC inflates this penalty to ensure that the probability of selecting the wrong model $\to 0$ as $n \to \infty$.

34.7 BIC selects the correct model

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

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

i.e. we select the wrong model if a χ^2_1 random variable exceeds $\log n$, and the probability of this $\to 0$ as $n \to \infty$ (for instance by applying Chebyshev).

It is also easy to verify that if the population mean was not 0, then as $n \to \infty$ BIC would correctly choose Model 2. More generally, BIC is known to be model selection consistent (see the original paper of Schwarz) in choosing between a fixed number of fixed-dimension models.

34.8 Hypothesis testing

Notice that one could also phrase model selection as a sequence of hypothesis testing problems. For our test problem it is natural to consider testing:

$$\begin{aligned} H_0 : &\quad \theta = 0 \\ H_1 : &\quad \theta \neq 0. \end{aligned}$$

In this case, we would reject the null if

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

and this would control 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.

34.9 Is it worth the fuss?

Notice that even when CV/AIC selects the wrong model, they are not off my much in terms of prediction error. In particular, even when we incorrectly select Model 2 in our example, we estimate the mean to be roughly $1/\sqrt{n}$ which is quite close to 0 (the true mean). In essence, we are choosing a parameter for Model 2 that is very close to Model 1.

From a practical standpoint however, this can affect interpretability, i.e. typically in linear regression for instance the model selected by CV will tend to have many small coefficients while the model selected by BIC will tend to be much more parsimonious/interpretable.