36-700: Probability and Mathematical Statistics I **Fall 2016**

Lecture 36: December 7

Lecturer: Siva Balakrishnan

36.1 Review and Outline

In the last class we were discussing the Metropolis-Hastings algorithm:

1. Detailed balance
2. The Metropolis-Hastings algorithm
3. The key point

Today we will briefly discuss some of the main ideas of model selection.

36.2 Introduction

We have previously seen the abstract setting (while discussing cross-validation): we have a collection of p models, $\mathcal{M} = \{M_1, M_2, \ldots, M_p\}$ and observe data $\{X_1, \ldots, X_n\}$ and want to choose a good model.

Here are some concrete examples:

Example 1: Suppose we have a response variable Y and a single covariate X, we would like to fit a polynomial regression function:

$$m(x) = \mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^p.$$

We need to choose the order of the polynomial. We can treat each order as a different model, and then we need to perform model selection.

Example 2: Suppose you would like to model a time series $Y_1, Y_2, \ldots$. A common model is an autoregressive model:

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

where the hypothesis is roughly that we can predict the current time-step well if we look k steps into the past. The model selection problem is then to pick a value for k.

Example 3: Suppose that you have many covariates $X_1, \ldots, X_d$ to use to predict a response Y. However, you do not want to use all the covariates (this will result in overfitting, and in a very complex model that is difficult to interpret). This is the problem of feature/variable selection. As a model selection problem: there are 2^d possible models (for each subset of features) and we would like to select one.

There are broadly a few different ways to do model selection:

1. Cross-validation (we have seen this one before)
2. Akaike Information Criterion (AIC) and closely related methods like Mallows C_p
3. Bayesian Information Criterion (BIC) and closely related methods like Minimum Description Length (MDL)

There are also different goals in model selection:

1. Find a model that fits the data best or predicts best,
2. Assume that one of the models truly generates the data, and to find the "true model" in this case.

In general, the first does not assume that you know anything about the data generating mechanism while the second makes a much more stringent assumption.

As a general guideline, AIC and cross-validation work for goal 1, while BIC works for goal 2. As we will see in a few minutes, AIC often selects "bigger" (more complex) models than BIC.

Our focus today will be on parametric models, here we have that each

$$M_j = \{p(x; \theta_j) : \theta_j \in \Theta_j\}.$$

36.3 Recap: Cross-Validation

In k-fold cross-validation we split the data into k folds, we fit each model (i.e. estimate $\widehat{\theta}_j$) on $k-1$ folds and evaluate its likelihood on the k-th fold, and then repeat this for each possible choice of the k-th fold and then average the results.

We have previously analyzed the train-validation split, which is a simplified version of cross-validation where we don't repeatedly average over the different folds.

Denote the fitted models $\{\widehat{\theta}_1, \ldots, \widehat{\theta}_p\}$. Define, the likelihood of $\widehat{\theta}$ as:

$$\mathcal{L}(\widehat{\theta}) = \mathbb{E} \log p_{\widehat{\theta}}(X).$$

In this case, we showed that we would select $\widehat{\theta}^*$ such that,

$$\mathcal{L}(\widehat{\theta}^*) \geq \mathcal{L}(\widehat{\theta}) - 2\sqrt{\frac{2B^2 \log(2p/\alpha)}{n}},$$

where $\widehat{\theta}$ is any other model in our set, and B is a bound on the log-likelihood. So we essentially select the best possible model in our set in terms of its likelihood.

In general, cross-validation is widely used and very powerful. Some drawbacks are that it does not use the entire data-set to train the model, and the second drawback is that it can be computationally intensive: requiring us to fit the p models, K times. The information criteria get around these drawbacks but are only guaranteed to work under stronger assumptions.

36.4 AIC

The first maybe natural question is why not use the likelihood directly to select a model, i.e. for each model i, we can evaluate:

$$\widehat{L}_i = \frac{1}{n} \sum_{i=1}^{n} \log p(X_i; \widehat{\theta}_i),$$

and then use the most likely model, i.e. pick $\widehat{\theta}^*$ that maximizes this expression.

The main issue is that this is a biased measure of the likelihood of a model, since the data is being used twice (once to fit the model and then once more to select a good model). Cross-validation fixed this by splitting the data.

Akaike instead suggested to calculate the bias and correct for it. The likelihood estimate above will usually be biased up. What he showed was that in parametric models the bias is approximately:

$$\text{bias}_i \approx \frac{d_i}{n},$$

where d_i was the number of parameters fit in the i-th model.

So the AIC rule is to pick the model that maximizes:

$$i^* = \arg\max_i \left(L_i - \frac{d_i}{n} \right).$$

It turns out that under a lot of conditions, this rule will select close to the most predictive model. Notice, that the entire data is used to fit the model.

36.5 BIC

In the setting where one model is assumed to be true and we would like to find it, a different rule is typically used. In this case, we select the model that maximizes:

$$i^* = \arg\max_i \left(L_i - \frac{d_i \log n}{2n} \right).$$

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

BIC is derived by assuming a prior on each of the models and an additional prior on the parameters of each model, and then making a long sequence of approximations to the posterior likelihood.

36.6 Model averaging

Rather than selecting a model, a different strategy that often has nice properties is to average the different models.

We will not discuss this in too much detail, but again suppose the goal is prediction (say in a regression problem). We might fit different regressors $\widehat{f}_1, \widehat{f}_2, \ldots, \widehat{f}_p$, and to predict the response at a new X we could use:

$$\widehat{Y} = \sum_{i=1}^{p} w_i \widehat{f}_i(X_i),$$

where the weights are positive and sum to 1. The intuition is just that if we select these weights cleverly we might be able to improve over any single model (in prediction). A common choice here is to select weights of the form:

$$w_i \propto \exp\left(L_i - \frac{d_i \log n}{2n} \right),$$

so we weight a model higher, if its BIC score is high.

Under many assumptions, one can show that this averaged prediction can be more accurate than any of the individual $\widehat{f}_i$.