STAT 535: Statistical Machine Learning **Autumn 2019**

Lecture 7: Model Selection and Prediction

Instructor: Yen-Chi Chen

Useful reference:

- Section 7 in *the elements of statistical learning* by Trever Hastie, Robert Tibshirani, Jerome Friedman.
- Section 13 (especially 13.6) in *all of statistics* by Larry Wasserman.
- Section 5.3 in *all of nonparametric statistics* by Larry Wasserman.

7.1 Introduction

Suppose that we observe $X_1, \cdots, X_n$ from an unknown distribution function. Under the assumption that our data comes from a parametric model p_θ with $\theta \in \Theta \subset \mathbb{R}^d$.

Let

$$\widehat{\theta}_n = \mathsf{argmax}_\theta L(\theta|X_1, \cdots, X_n) = \mathsf{argmax}_\theta \ell(\theta|X_1, \cdots, X_n)$$

be the MLE. For simplicity, we denote

$$\ell_n = \ell(\widehat{\theta}_n|X_1, \cdots, X_n)$$

as the maximal value of the empirical log-likelihood function and we define

$$\bar{\ell}(\theta) = \mathbb{E}(\ell(\theta|X_1))$$

as the population log-likelihood function. The population likelihood function implies the population MLE, which is

$$\theta^* = \mathsf{argmax}_\theta \bar{\ell}(\theta).$$

7.2 AIC: Akaike information criterion

The AIC is an information criterion that is common used for model selection. In model selection, the AIC propose the following criterion:

$$AIC = 2d - 2\ell_n,$$

where d is the dimension of the model.

The idea of AIC is to adjust the empirical risk to be an unbiased estimator of the true risk in a parametric model. Under a likelihood framework, the loss function is the negative log-likelihood function so the empirical risk is

$$\widehat{R}_n(\widehat{\theta}_n) = -\ell_n = -\ell(\widehat{\theta}_n|X_1, \cdots, X_n) = -\ell_n(\widehat{\theta}_n).$$

On the other hand, the true risk of the MLE is

$$R(\widehat{\theta}_n) = \mathbb{E}(-n\bar{\ell}(\widehat{\theta}_n)).$$

Note that we multiply it by n to reflect the fact that in the empirical risk, we did not divide it by n.

To derive the AIC, we examine the asymptotic bias of the empirical risk $\widehat{R}_n(\widehat{\theta}_n)$ versus the true risk $R(\widehat{\theta}_n)$.

Analysis of true risk. To analyze the true risk $R(\widehat{\theta}_n)$, we first investigate the asymptotic behavior of $\bar{\ell}(\widehat{\theta}_n)$ around θ^*:

$$\begin{aligned}\bar{\ell}(\widehat{\theta}_n) &\approx \bar{\ell}(\theta^*) + (\widehat{\theta}_n - \theta^*)^T \underbrace{\nabla\bar{\ell}(\theta^*)}_{=0} + \frac{1}{2}(\widehat{\theta}_n - \theta^*)^T \underbrace{\nabla\nabla\bar{\ell}(\theta^*)}_{=I(\theta^*)}(\widehat{\theta}_n - \theta^*)\\ &= \bar{\ell}(\theta^*) + \frac{1}{2}(\widehat{\theta}_n - \theta^*)^T I(\theta^*)(\widehat{\theta}_n - \theta^*).\end{aligned}$$

Thus, the true risk is

$$R(\widehat{\theta}_n) = -n\mathbb{E}(\bar{\ell}(\widehat{\theta}_n)) \approx -n\bar{\ell}(\theta^*) - \frac{n}{2}\mathbb{E}\left((\widehat{\theta}_n - \theta^*)^T I(\theta^*)(\widehat{\theta}_n - \theta^*)\right). \tag{7.1}$$

Analysis of expected empirical risk. For the expected empirical risk, we first expand ℓ_n as follows:

$$\begin{aligned}\ell_n &= \sum_{i=1}^n \ell(\widehat{\theta}_n|X_i)\\ &\approx \sum_{i=1}^n \ell(\theta^*|X_i) + \underbrace{(\widehat{\theta}_n - \theta^*)^T \sum_{i=1}^n \nabla\ell(\theta^*|X_i)}_{(I)} + \underbrace{\frac{1}{2}(\widehat{\theta}_n - \theta^*)^T \sum_{i=1}^n \nabla\nabla\ell(\theta^*|X_i)(\widehat{\theta}_n - \theta^*)}_{=(II)}.\end{aligned} \tag{7.2}$$

The expectation of the first quantity is $\bar{\ell}(\theta^*)$, which agrees with the first term in the true risk so all we need is to understand the behavior of the rest two quantities.

For the first quantity, using the fact that $\sum_{i=1}^n \nabla\ell(\widehat{\theta}_n|X_i) = 0$,

$$\begin{aligned}\sum_{i=1}^n \nabla\ell(\theta^*|X_i) &= \sum_{i=1}^n \nabla(\ell(\theta^*|X_i) - \ell(\widehat{\theta}_n|X_i))\\ &\approx \left(\sum_{i=1}^n \nabla\nabla\ell(\theta^*|X_i)\right)(\theta^* - \widehat{\theta}_n)\\ &\approx \underbrace{(\nabla\nabla\mathbb{E}(\ell(\theta^*|X_i)))}_{=I(\theta^*)} n(\theta^* - \widehat{\theta}_n).\end{aligned}$$

Thus,

$$(I) \approx -n(\widehat{\theta}_n - \theta^*)^T I(\theta^*)(\widehat{\theta}_n - \theta^*).$$

For quantity (II), note that

$$\frac{1}{n}\sum_{i=1}^n \nabla\nabla\ell(\theta^*|X_i) \approx \nabla\nabla\mathbb{E}(\ell(\theta^*|X_i)) = I(\theta^*)$$

so

$$\begin{aligned}(II) &= \frac{1}{2}(\widehat{\theta}_n - \theta^*)^T \sum_{i=1}^n \nabla\nabla\ell(\theta^*|X_i)(\widehat{\theta}_n - \theta^*)\\ &= \frac{n}{2}(\widehat{\theta}_n - \theta^*)^T I(\theta^*)(\widehat{\theta}_n - \theta^*)\end{aligned}$$

Combining (I) and (II) into equation (7.2) and taking the expectation, we obtain

$$\mathbb{E}(\widehat{R}_n(\widehat{\theta}_n)) = -\mathbb{E}(\ell_n) = -n\bar{\ell}(\theta^*) + \frac{n}{2}\mathbb{E}\left((\widehat{\theta}_n - \theta^*)^T I(\theta^*)(\widehat{\theta}_n - \theta^*)\right)$$

Comparing this to equation (7.1), we obtain

$$\mathbb{E}(\widehat{R}_n(\widehat{\theta}_n)) - R(\widehat{\theta}_n) = -n\mathbb{E}\left((\widehat{\theta}_n - \theta^*)^T I(\theta^*)(\widehat{\theta}_n - \theta^*)\right).$$

Note that by the theory of MLE, one can easily shown that

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) \approx N(0, I^{-1}(\theta^*))$$

so

$$n(\widehat{\theta}_n - \theta^*)^T I(\theta^*)(\widehat{\theta}_n - \theta^*) \approx \chi^2_d,$$

which implies that[1]

$$n\mathbb{E}\left((\widehat{\theta}_n - \theta^*)^T I(\theta^*)(\widehat{\theta}_n - \theta^*)\right) = d.$$

Thus, to make sure that we have an asymptotic unbiased estimator of the true risk $R(\widehat{\theta}_n)$, we need to modify the empirical risk by

$$\widehat{R}_n(\widehat{\theta}_n) + d = -\ell_n + d.$$

Multiplying this quantity by 2, we obtain the AIC

$$AIC = 2d - 2\ell_n.$$

Remark. From the derivation of AIC, we see that the goal of the AIC is to adjust the model so that we are comparing unbiased estimates of the true risks across different models. Thus, the model selected by minimizing the AIC can be viewed as the model selected by minimizing unbiased estimates of the true risks. From the risk minimization point of view, this is trying to make a prediction using a good risk estimator. Thus, some people would common that the design of AIC is to choose a model that makes good predictions.

7.3 BIC: Bayesian information criterion

Another common approach for model selection is the BIC:

$$BIC = d\log n - 2\ell_n,$$

where again d denotes the dimension of the model.

Here is the derivation of the BIC. In the Bayesian setting, we place a prior $\pi(m)$ over all possible models and within each model, we place a prior over parameters $p(\theta|m)$. The BIC is a Bayesian criterion, which means that we will select model according to the posterior distribution of each model m. Namely, we will try to derive $\pi(m|X_1, \cdots, X_n)$.

By Bayes rule, we have

$$\pi(m|X_1, \cdots, X_n) = \frac{\pi(m, X_1, \cdots, X_n)}{p(X_1, \cdots, X_n)} \propto p(X_1, \cdots, X_n|m)\pi(m)$$

so all we need is to derive the marginal density in a model $p(X_1, \cdots, X_n|m)$.

[1] Formally, we need a few more conditions; the convergence in distribution is not enough for the convergence in expectation.

With a prior $\pi(\theta|m)$, this marginal density can be written as

$$p(X_1, \cdots, X_n|m) = \int p(X_1, \cdots, X_n|\theta, m)\pi(\theta|m)d\theta. \tag{7.3}$$

Suppose that the model m is correct in the sense that under model m, there exists $\theta^* \in \Theta$ such that the data are indeed generated from $p(x|\theta^*)$.

Using the log-likelihood function, we can expand

$$p(X_1, \cdots, X_n|\theta, m) = e^{\ell(\theta|X_1,\cdots,X_n,m)} = e^{\sum_{i=1}^n \ell(\theta|X_i,m)}. \tag{7.4}$$

Asymptotically, the log-likelihood function can be further expand

$$\begin{aligned}\sum_{i=1}^n \ell(\theta|X_i, m) = \sum_{i=1}^n \ell(\theta^*|X_i, m) &+ (\theta - \theta^*)^T \underbrace{\sum_{i=1}^n \nabla\ell(\theta^*|X_i, m)}_{=0} \\ &+ (\theta - \theta^*)^T \sum_{i=1}^n \nabla\nabla\ell(\theta^*|X_i, m)(\theta - \theta^*) + \text{ small terms} \\ &= \ell_n + n(\theta - \theta^*)^T \underbrace{\frac{1}{n}\sum_{i=1}^n \nabla\nabla\ell(\theta^*|X_i, m)}_{\approx -I(\theta^*)}(\theta - \theta^*) + \text{ small terms,}\end{aligned}$$

where $I(\theta^*)$ is the Fisher's information matrix. Plugging this into equation (7.4) and ignore the reminder terms, we obtain

$$p(X_1, \cdots, X_n|\theta, m) \approx e^{\ell_n - n(\theta-\theta^*)^T I(\theta^*)(\theta-\theta^*)}. \tag{7.5}$$

Thus, equation (7.3) can be rewritten as

$$p(X_1, \cdots, X_n|m) \approx e^{\ell_n} \int e^{-n(\theta-\theta^*)^T I(\theta^*)(\theta-\theta^*)}\pi(\theta|m)d\theta. \tag{7.6}$$

To compute the above integral, consider a random vector $Y \sim N(0, \frac{1}{n}I(\theta^*))$. The expectation

$$\begin{aligned}\mathbb{E}(\pi(Y|m)) &= \left(\frac{n}{2\pi}\right)^{d/2} \mathsf{det}^{-1}(I(\theta^*)) \int e^{-n(y-\theta^*)^T I(\theta^*)(y-\theta^*)}\pi(y|m)dy \\ &\approx \pi(\widehat{\theta}_n|m)\end{aligned}$$

when $n \to \infty$. This implies that

$$\int e^{-n(\theta-\theta^*)^T I(\theta^*)(\theta-\theta^*)}\pi(\theta|m)d\theta \approx \left(\frac{2\pi}{n}\right)^{d/2} \mathsf{det}(I(\theta^*))\pi(\widehat{\theta}_n|m).$$

Putting this into equation (7.6), we conclude that the Bayesian evidence

$$p(X_1, \cdots, X_n|m) \approx e^{\ell_n}\left(\frac{2\pi}{n}\right)^{d/2} \mathsf{det}(I(\theta^*))\pi(\widehat{\theta}_n|m)$$

so the log evidence is

$$\log p(X_1, \cdots, X_n|m) \approx \ell_n - \frac{d}{2}\log n + \frac{d}{2}\log(2\pi) + \log \mathsf{det}(I(\theta^*)) + \log \pi(\widehat{\theta}_n|m).$$

The only quantity that would increases with respect to the sample size n are the first two quantities so after multiplying by -2 and keeping only the dominating two terms, we obtain

$$BIC = d\log n - 2\ell_n.$$

Remark. Although the BIC leads to a criterion similar to the AIC, the reasoning is somewhat different. In the construction of BIC, the effect of priors are ignored since we are working on the limiting regime but we still use the Bayesian evidence as a model selection criterion. We are selecting the model with the highest evidence. When the data is indeed generated from one of the model in the collection of models we are choosing from, the posterior will concentrate on this correct model. So BIC would eventually be able to select this model. Therefore, some people would argue that unlike AIC that chooses the best predictive model, the BIC attempts to select the true model if it exists in the model set.

7.4 Information criterion in regression

Using information criterion in a regression problem is very common in practice. However, there is a caveat in the construction. The information criterion we derived is based on a likelihood framework but the regression problem is often formulated as an empirical risk minimization process, e.g., least squared approach. So we need to associate the likelihood function to the loss function used in the regression problem to properly construct a model selection criterion. Here we gave one example of using the least square approach under a linear regression model.

Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be the observed data such that $X_i \in \mathbb{R}^d$ and $Y_i \in R$. A regression model associate X_i and Y_i via

$$Y_i = X_i^T\beta + \epsilon_i,$$

where ϵ_i's are the noise. To associate the least square method to a likelihood framework, we assume that $\epsilon_i \sim N(0, \sigma^2)$. Under this framework, one can easily shown that the MLE of β and the least squared estimate are the same.

What will the likelihood function be like in this case? For simplicity, we consider the fixed design such that X_i's are non-random and the only random quantity is the noise ϵ_i's. Because $\epsilon_i = Y_i - X_i^T\beta$, the log-likelihood function will be

$$\ell(\beta, \sigma^2|X_i, Y_i) = \log p_\epsilon(Y_i - X_i^T\beta; \beta, \sigma^2) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{\sigma^2}(Y_i - X_i^T\beta)^2.$$

Let $\widehat{\beta}$ be the least square estimate (MLE) and $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n e_i^2$ be the MLE of the noise level and $e_i = Y_i - X_i^T\widehat{\beta}$ is the residual.

The empirical risk under the MLE/least square estimate is

$$\begin{aligned}
\ell_n &= \sum_{i=1}^n \ell(\beta, \sigma^2 | X_i, Y_i) \\
&= -\frac{n}{2}\log(2\pi\widehat{\sigma}^2) - \frac{1}{\widehat{\sigma}^2}\underbrace{\sum_{i=1}^n (Y_i - X_i^T\beta)^2}_{=n\widehat{\sigma}^2} \\
&= -\frac{n}{2}(\log 2 + \log\pi + \log\widehat{\sigma}^2) - n \\
&= C_n - \frac{n}{2}\log\left(\frac{1}{n}\sum_{i=1}^n e_i^2\right) \\
&= C_n - \frac{n}{2}\log\left(\frac{1}{n}\mathsf{RSS}_n\right),
\end{aligned}$$

where $\mathsf{RSS}_n = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - X_i^T\beta)^2$ is the least square objective function or the so-called residual sum of squares.

The quantity C_n is invariant across models/variables we choose. Thus, the AIC and BIC of the regression model will be

$$\begin{aligned}
AIC &= 2d - n\log\left(\frac{1}{n}\mathsf{RSS}_n\right) \\
BIC &= d\log n - n\log\left(\frac{1}{n}\mathsf{RSS}_n\right).
\end{aligned}$$

7.5 Prediction risk in regression

Now we consider a general regression prediction problem where we observe $(X_1, Y_1), \cdots, (X_n, Y_n)$ where $X_i \in \mathbb{R}^d$ and $Y_i \in R$ and they are from an unknown joint distribution $F_{X,Y}$. Here we do not specify a particular model of the regression function and all we assume is

$$Y_i = m(X_i) + \epsilon_i,$$

where ϵ_i's are mean 0 noise with $\mathsf{Var}(\epsilon_i | X_i = x) = \sigma^2(x)$ and $m(x) = \mathbb{E}(Y|X = x)$ is the regression function. Suppose that we have an estimator $\widehat{m}$ of m and we would like to know the prediction risk (expected loss in predicting a future outcome) of this estimator.

To define the prediction risk, let $X_{\mathsf{new}}, Y_{\mathsf{new}}$ be a new pair of observation from $F_{X,Y}$. Then the prediction risk is defined as

$$R(\widehat{m}) = \mathbb{E}((\widehat{Y}_{\mathsf{new}} - Y_{\mathsf{new}})^2) = \mathbb{E}((\widehat{m}(X_{\mathsf{new}}) - Y_{\mathsf{new}})^2),$$

where $\widehat{Y} = \widehat{m}(X)$ is the predictive value of Y given covariate X.

Let $\bar{m}(x) = \mathbb{E}(\widehat{m}(x))$ be the expected regression function of the estimator/predictor $\widehat{m}$. We can decompose the prediction risk using the following expansion:

$$\begin{aligned}
R(\widehat{m}) &= \mathbb{E}((\widehat{m}(X_{\mathsf{new}}) - Y_{\mathsf{new}})^2) \\
&= \mathbb{E}(\mathbb{E}((\widehat{m}(X_{\mathsf{new}}) - Y_{\mathsf{new}})^2 | X_{\mathsf{new}})) \\
&= \mathbb{E}(R(X_{\mathsf{new}})) \\
&= \int R(x) p_X(x) dx
\end{aligned}$$

where

$$R(x) = \mathbb{E}((\widehat{m}(X_{\mathsf{new}}) - Y_{\mathsf{new}})^2 | X_{\mathsf{new}} = x)$$

is a local predictive risk and p_X is the PDF of the covariate X.

The local predictive risk can be decomposed as

$$\begin{aligned} R(x) &= \mathbb{E}((\widehat{m}(X_{\mathsf{new}}) - Y_{\mathsf{new}})^2 | X_{\mathsf{new}} = x) \\ &= \mathbb{E}\left((\underbrace{\widehat{m}(X_{\mathsf{new}}) - \bar{m}(X_{\mathsf{new}})}_{\text{variance of } \widehat{m}} + \underbrace{\bar{m}(X_{\mathsf{new}}) - m(X_{\mathsf{new}})}_{\text{bias of } \widehat{m}} + \underbrace{m(X_{\mathsf{new}}) - Y_{\mathsf{new}}}_{\text{intrinsic variance}})^2 | X_{\mathsf{new}} = x \right) \\ &= \mathbb{E}(V(X_{\mathsf{new}}) + b^2(X_{\mathsf{new}}) + \sigma^2(X_{\mathsf{new}}) | X_{\mathsf{new}} = x), \end{aligned}$$

where

$$\begin{aligned} V(x) &= \mathsf{Var}(\widehat{m}(X_{\mathsf{new}}) | X_{\mathsf{new}} = x) \\ b(x) &= \bar{m}(x) - m(x). \end{aligned}$$

Thus, the predictive risk can be decomposed into

$$R(\widehat{m}) = \underbrace{\mathbb{E}(b^2(X_{\mathsf{new}}))}_{\text{bias}} + \underbrace{\mathbb{E}(V(X_{\mathsf{new}}))}_{\text{Variance}} + \underbrace{\mathbb{E}(\sigma^2(X_{\mathsf{new}}))}_{\text{intrinsic error}}.$$

To compare the performance of prediction, we should use a good estimate of the predictive risk. A naive estimator is the training error (or empirical risk), which is

$$\widehat{R}(\widehat{m}) = \frac{1}{n}\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \widehat{m}(X_i))^2.$$

Although this estimator seems to be intuitively correct, it may suffer from a severe bias! To see this, note that the expected value of the empirical risk is

$$\mathbb{E}(\widehat{R}(\widehat{m})) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(Y_i - \widehat{m}(X_i))^2.$$

For each i,

$$\begin{aligned} \mathbb{E}(Y_i - \widehat{m}(X_i))^2 &= \mathbb{E}(Y_i - m(X_i) + m(X_i) - \bar{m}(X_i) + \bar{m}(X_i) - \widehat{m}(X_i))^2 \\ &= \mathbb{E}(\sigma^2(X_i)) + \mathbb{E}(b^2(X_i)) + \mathbb{E}(V(X_i)) + 2\mathbb{E}((Y_i - m(X_i))(\bar{m}(X_i) - \widehat{m}(X_i))) \\ &= R(\widehat{m}) - 2\mathbb{E}((Y_i - m(X_i))(\widehat{Y}_i - \bar{m}(X_i))) \\ &= R(\widehat{m}) - 2\mathbb{E}(\mathsf{Cov}(Y_i, \widehat{Y}_i | X_i)). \end{aligned}$$

Thus, the empirical risk and the predictive risk has the follow relationship

$$E(\widehat{R}(\widehat{m})) = R(\widehat{m}) - \frac{2}{n}\sum_{i=1}^n \mathbb{E}(\mathsf{Cov}(Y_i, \widehat{Y}_i | X_i)),$$

which implies that the empirical risk is always underestimate the predictive risk. This is especially more severe when the predictor $\widehat{Y}_i$ is highly correlated with the i-th observation– a common scenario when we are overfitting the model.

Note that in the fix design case (covariate X are non-random), the above expression reduces to

$$E(\widehat{R}(\widehat{m})) = R(\widehat{m}) - \frac{2}{n}\sum_{i=1}^{n} \mathsf{Cov}(Y_i, \widehat{Y}_i),$$

and the quantity $\mathsf{Cov}(Y_i, \widehat{Y}_i)$ is called the *degrees of freedom.* For more details, see the following papers:

1. Tibshirani, Ryan J. "Degrees of freedom and model search." *Statistica Sinica* (2015): 1265-1296.
2. Tibshirani, Ryan J., and Saharon Rosset. "Excess Optimism: How Biased is the Apparent Error of an Estimator Tuned by SURE?." *Journal of the American Statistical Association* 114, no. 526 (2019): 697-712.

For a general regression model, a simple and consistent estimate of the predictive risk is the *cross-validation (CV).* With the growing power of computing, it is a very convenient tool nowadays.

If we are using a linear smoother, there is a closed-form of the leave-one-out CV:

$$\widehat{R}_{\mathsf{LOO}}(\widehat{m}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \widehat{m}(X_i)}{1 - L_{ii}}\right),$$

where L_{ij} is the (i, j) component in the linear smoother matrix L, i.e., $\widehat{\mathbf{Y}} = L\mathbf{Y}$. Also, there is a generalized method called the *generalized cross-validation (GCV)* that is available for obtaining a good estimate of the predictive risk:

$$\widehat{R}_{\mathsf{GCV}}(\widehat{m}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \widehat{m}(X_i)}{1 - \nu/n}\right),$$

where $\nu = \sum_i L_{ii}$.

Remark (Mallow's Cp). The Mallow's Cp is an approach to estimate $\mathsf{Cov}(Y_i, \widehat{Y}_i)$ and correct the empirical risk to obtain a better estimate of the predictive risk. The Mallow's Cp is

$$\widehat{R}_{\mathsf{Cp}} = \widehat{R}(\widehat{m}) + \frac{2d\widehat{\sigma}^2}{n},$$

where $\widehat{\sigma}^2$ is an estimate of the overall noise level $\sigma^2 = \int \sigma^2(x)p_X(x)dx$ and d is the dimension of the model. In the case of linear smoother, $d = \nu$. One can simply use the average of squared residuals in this case. The Mallow's Cp also leads to a model selection criterion but the goodness-of-fit part is different from AIC/BIC.