Comparing Models

Prof. Dennis Sun

Data 401
Overfitting

As you add more variables to a model, the fit only improves.

Consider fitting different polynomial models to the following data:

Intuitively, $p = 2$ is the best, even though MSE is lower for $p = 4$. 
What would a statistician say?

How might a statistician explain why $p = 2$ is a better model than $p = 1$, but $p = 4$ is worse than $p = 2$?

**Principle of Parsimony:** If two models fit the data equally well, then the simpler model is preferred.
What would a computer scientist say?

How might a computer scientist explain why $p = 2$ is a better model than $p = 1$, but $p = 4$ is worse than $p = 2$?

**Prediction Error:** Although the $p = 4$ model fits the current data better, if you use it to predict on new data, it will perform worse because of **overfitting**.
**AIC**

The *Akaike Information Criterion* (AIC) chooses the model with the lowest value of

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip})^2 + 2p.$$  

The first term can only go down as we increase the number of variables.

The second term is a penalty for the number of variables $p$. So if two models have similar MSEs, AIC will prefer the one with fewer variables.
The **Bayesian Information Criterion** (BIC) chooses the model with the lowest value of

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip})^2 + p \log n.$$ 

The only difference between AIC and BIC is in the penalty for the number of variables. The penalty of $2p$ is replaced by $p \log(n)$.

Which one will favor parsimonious models more strongly: AIC or BIC? **BIC**
Other Criteria

- Adjusted $R^2$
- Partial $F$-tests allow us to compare two nested models:

\[
Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i
\]

\[
Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i
\]

The first is just a special case of the second, where $\beta_3 = \beta_4 = 0$.

The partial $F$-test tests the hypothesis

\[
H_0 : \beta_3 = \beta_4 = 0.
\]

If we fail to reject the null hypothesis, then we should not add the additional variables $x_{i3}$ and $x_{i4}$ because they do not improve the fit enough to justify the complexity.
• Instead of penalizing the number of parameters, we can also optimize for **prediction error**, the error if we were to apply this model to future data.

• How do we estimate prediction error?

**Answer:** Split the data into training and test sets. Use only the training data to fit the model. Then, we can evaluate the error of our predictions on the test set. This is called the **test error**.
Cross Validation

- You might be concerned that our estimate of the prediction error was based on a single test set.
- Cross validation creates several test sets, as follows:
  1. First, the data is divided into \( k \) folds.
  2. One at a time, each fold is used as the test set and the model is trained on the remaining data. Then, the test error is calculated using the held-out fold.
  3. In the end, we will get \( k \) estimates of the prediction error.

Example of 4-fold cross validation

\[
X = \begin{array}{c}
\text{TEST} \\
\text{TEST} \\
\text{TEST} \\
\text{TEST}
\end{array}
\]