Model Assessment and Selection

36-290 – Statistical Research Methodology

Week 4 Thursday – Fall 2021

What's the Difference?

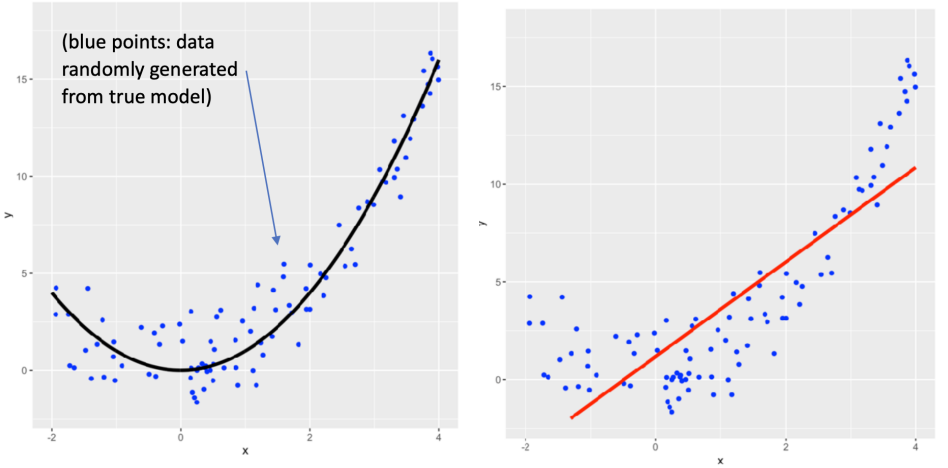
Model Assessment:

- evaluating how well a learned model performs, via the use of a single-number metric

Model Selection:

- selecting the best model from a suite of learned models (e.g., linear regression, random forest, etc.)

Model "Flexibility"



The data (blue points) are generated from a smoothly varying non-linear model (black line), with random noise added:

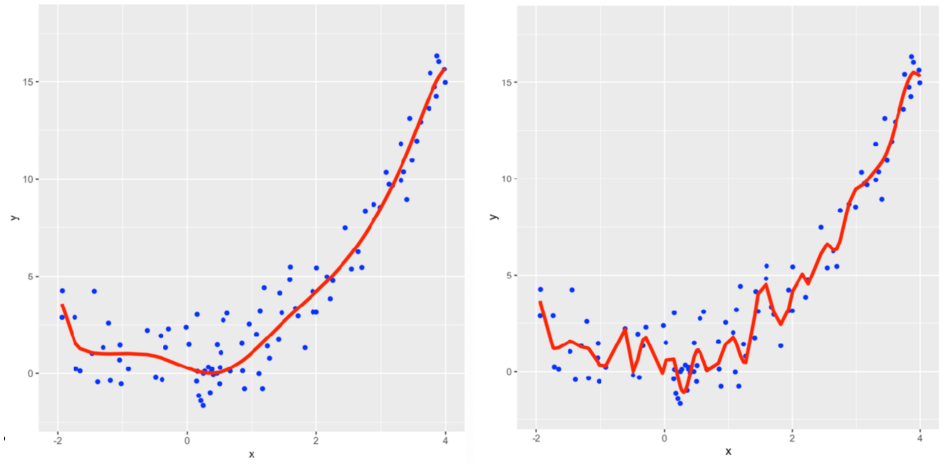
$$Y = f(X) + \epsilon$$

In the right panel, the red line shows a simple linear regression fit to the data: linear regression involves the use of an inflexible, fully parametrized model, and thus it can neither provide a good estimate of $f(X)$ nor can it "overfit" by modeling the noisy deviations of the data from $f(X)$.

If data were repeatedly generated from the true model and fit via linear regression, we'd see that

- The *average* estimated y value is offset from the truth. This is "high bias."
- The variance in the estimated y values is relatively small. This is "low variance."

Model "Flexibility"



The right panel above shows a model that is overly flexible: it can provide a good estimate of $f(X)$ but it goes too far and overfits by modeling the noisy deviations from $f(X)$. Such a model is not "generalizable": it will tend to do a bad job of predicting the response given a new predictor X_o that was not used in learning the model.

The left panel shows a model that is close to being the optimal model: neither too flexible nor inflexible.

If data were repeatedly generated from the true model and fit via the model at right, we'd see that

- The *average* estimated y value approximately matches the truth. This is "low bias."
- The variance in the estimated y values is relatively large. This is "high variance."

How to Deal with Flexibility

So...we want to learn a statistical model that provides a good estimate of $f(X)$ without overfitting.

There are two common approaches:

- We can split the data into two groups: one used to train models, and another used to test them. By assessing models using "held-out" test set data, we act to ensure that we get a generalizable(!) estimate of $f(X)$.

- We can repeat data splitting k times, with each datum being placed in the "held-out" group exactly once. This is *k-fold cross validation*. The general heuristic is that k should be 5 or 10.

k-fold cross validation is the preferred approach, but the tradeoff is that CV analyses take $\sim k$ times longer than analyses that utilize data splitting.

Model Assessment: Mean-Squared Error

When the response variable is quantitative (like it is in the example in these notes), the most commonly used assessment metric is the mean-squared error, or MSE, computed with the *test-set data* only:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \,.$$

Y_i is the value of the response for the i^{th} test-set datum and $\hat{Y}_i$ is the predicted response value for the i^{th} test-set datum.

It can be shown that:

$$\text{MSE} = (\text{Bias})^2 + \text{Variance}\,.$$

Thus an overly inflexible model, a high-bias/low-variance model, will have a high MSE value, as will an overly flexible model (low-bias but high-variance). The best regression model is that which balances bias and variance!

Let's step back for a moment. A *loss function*, also known as a *cost function*, is a metric that represents the quality of fit of a model. There is no unique loss function, but the most commonly used one is the quadratic loss function, based on the squared difference between model predictions and observed data. Because data are noisy, possible values of the loss follow a distribution...and the mean, or expected value, of this distribution is dubbed the *risk*.

The MSE is an estimator of the risk for the quadratic loss function.

Model Assessment: What About Classification?

The MSE is an assessment metric that is appropriate to use when the response variable is quantitative.

What about if the response variable is categorical? The choice of metric is unfortunately not quite so clear-cut. Common ones include the *misclassification rate* or *MCR* (what percentage of predictions are wrong) and the *area under curve*. And note that interpretation can be affected by *class imbalance*: if two classes are equally represented in a dataset, an MCR of 2% is quite good; but if one class comprises 99% of the data, that 2% is no longer such a good result.

We will return to discuss classification metrics in a future lecture.

Reproducibility

An important aspect of a statistical analysis is that it be reproducible. You should...

1. Record your analysis in a notebook, via, e.g., `R Markdown` or `Jupyter`. A notebook should be complete such that if you give it and datasets to someone, that someone should be able to recreate the entire analysis and achieve the exact same results. To ensure the achivement of the exact same results, you should...

2. Manually set the random-number generator seed before each instance of random sampling in your analysis (such as when you assign data to training or test sets, or to folds):

```
set.seed(101)     # can be any number...
sample(10,3)      # sample three numbers between 1 and 10 inclusive
```

```
## [1]  9 10  6
```

```
set.seed(101)
sample(10,3)      # voila: the same three numbers!
```

```
## [1]  9 10  6
```

Model Selection

As mentioned on the first slide, model selection is picking the best model from a suite of possible models. This can be as simple as picking the regression model with the best *MSE* or the classification model with the best *MCR*. However, two things must be kept in mind:

1. To ensure an apples-to-apples comparison of metrics, every model should be learned using *the same training and test set data*! Do not resample the data between the time when you, e.g., perform linear regression and when you perform random forest.

2. An assessment metric is a *random variable*, i.e., if you choose different data to be in your training set, the metric will be different.

For regression, a third point should be kept in mind:

1. A metric like the MSE is *unit-dependent*: an MSE of 0.001 in one analysis context is not necessarily better or worse than an MSE of 100 in a completely different analysis context.