Introduction to Machine Learning

Session 1c: Assessing Model Accuracy

Reto Wüest

Department of Political Science and International Relations
University of Geneva

Outline

Selection of a Machine Learning Method

No-Free-Lunch Theorem

There is no universal learning method that performs best on all learning tasks.

This implies that...

- We need to decide for any given data set which method performs best.
- To evaluate the performance of a method on a data set, we need a way to measure how well its predictions match the observed data.

Assessing Model Accuracy in Regression Problems

Measuring the Quality of Fit of a Method

- In regression problems, the most commonly used performance measure is the mean squared error (MSE)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(\mathbf{x}_i)\right)^2, \tag{1}$$

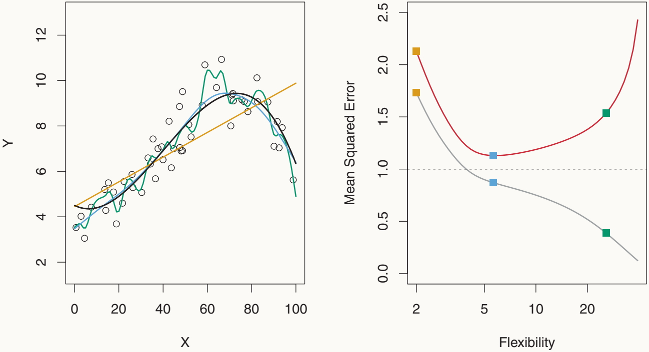
where $\hat{f}(\mathbf{x}_i)$ is the prediction that $\hat{f}$ produces for the ith observation.

- The MSE in (1) is computed using the training data, so it is the training MSE.
- However, what we care about is how well the method performs on new (i.e., previously unseen) test data $\{(\widetilde{\mathbf{x}}_i, \widetilde{y}_i)\}_{i=1,\ldots,m}$.
- We therefore select the method that minimizes the test MSE

$$\text{test MSE} = \frac{1}{m}\sum_{i=1}^{m}\left(\widetilde{y}_i - \hat{f}(\widetilde{\mathbf{x}}_i)\right)^2. \tag{2}$$

Measuring the Quality of Fit of a Method

- What happens if we select instead the method that minimizes the training MSE in (1)?



(Source: James et al. 2013, 31)

- Overfitting the data: a model that is less flexible than the one we selected would have yielded a smaller test MSE.

- The U-shape in the test MSE curve is the result of two competing properties of learning methods.
- The expected test MSE for value x_0 can be decomposed into the sum of three quantities

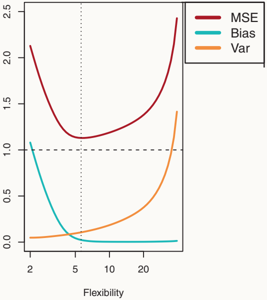
$$E\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = Var\left[\hat{f}(x_0)\right] + \left(\text{Bias}\left[\hat{f}(x_0)\right]\right)^2 + \underbrace{Var\left[\varepsilon\right]}_{\substack{\text{Irreducible}\\ \text{error}}}. \tag{3}$$

- To minimize the expected test MSE, we need to select a method that simultaneously achieves low variance and low bias.

The Bias-Variance Trade-Off

- What are the bias and variance of a method?
- **Bias:** The error that we introduce by approximating the true f by the estimate $\hat{f}$.
- **Variance:** Different training data sets result in a different $\hat{f}$. The variance refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set.

- More flexible methods have higher variance, while less flexible methods have higher bias. This is the bias-variance trade-off.



(Source: James et al. 2013, 36)

- In practice f is unobserved, making it impossible to explicitly compute the test MSE, bias, or variance for a method.
- We need to estimate the test MSE based on training data (e.g., by using cross-validation).

Cross-Validation

- Cross-validation (CV) is a re-sampling method that can be used to estimate the test error of a learning method based on the training data.
- Randomly split the n training observations into $2 \leq k \leq n$ non-overlapping groups (folds) of approximately equal size.
- Use the first fold as the validation data set and the remaining folds as the training data set.
- Fit the model on the training observations.
- Use the fitted model to make predictions for the excluded observations and compute the MSE.

Cross-Validation

- Repeat the procedure, each time using another fold as the validation data set. This gives k estimates of the test error, $\mathsf{MSE}_1, \mathsf{MSE}_2, \ldots, \mathsf{MSE}_k$.
- The CV estimate for the test MSE is given by the average

$$\mathsf{CV}_{(k)} = \frac{1}{k} \sum_{i=1}^{k} \mathsf{MSE}_i. \tag{4}$$

- If $k < n$, then this procedure is called k-fold cross-validation.
- If $k = n$, then we call it leave-one-out cross-validation (LOOCV).

Assessing Model Accuracy in Classification Problems

Measuring the Error Rate of a Method

- Suppose that we estimate f on the basis of training data $\{(\mathbf{x}_i, y_i)\}_{i=1,\dots,n}$, where $y_1, \dots, y_n$ are qualitative.
- The most commonly used approach for quantifying the accuracy of $\hat{f}$ is the error rate

$$\text{error rate} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(y_i \neq \widehat{y}_i), \tag{5}$$

where $\widehat{y}_i$ is the predicted class label for i using $\hat{f}$ and $\mathbb{1}(y_i \neq \widehat{y}_i)$ is an indicator variable that equals 1 if $y_i \neq \widehat{y}_i$ (misclassification) and 0 if $y_i = \hat{y}_i$ (correct classification).

- The error rate in (5) is the training error rate because it is computed based on the training data.

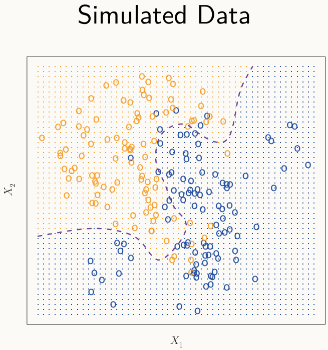
Measuring the Error Rate of a Method

- Again, however, we are more interested in selecting a method that minimizes the error rate on new test data $\{(\widetilde{\mathbf{x}}_i, \widetilde{y}_i)\}_{i=1,\dots,m}$

$$\text{test error rate} = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}(\widetilde{y}_i \neq \widehat{\widetilde{y}}_i). \tag{6}$$

- One can show that the test error rate is minimized by the Bayes classifier, which assigns each observation to the most likely class, given its predictor values.
- The Bayes classifier produces the lowest possible test error rate (the Bayes error rate).
- The Bayes error rate is analogous to the irreducible error in the regression setting.

Measuring the Error Rate of a Method



(Source: James et al. 2013, 38)

For each $X = x$, there is a probability that Y is orange or blue. The orange region is the set of x for which $\Pr(Y = \text{orange} \mid X = x) > 0.5$ and the blue region is the set for which $\Pr(Y = \text{orange} \mid X = x) \leq 0.5$. The dashed line is the Bayes decision boundary.

Measuring the Error Rate of a Method

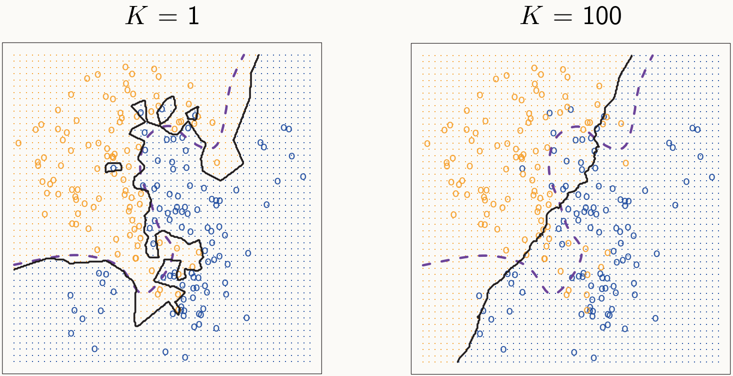
- For real data, we do not know $\Pr(Y = j \mid X = x)$, so we cannot compute the Bayes classifier.
- We need to estimate $\Pr(Y \mid X)$ and then classify a given observation to the class with the highest estimated probability.
- One method to do so is the K-nearest neighbors (KNN) classifier. Given a $K \in \mathbb{Z}_{>0}$ and a test observation x_0, KNN identifies the K points in the training data closest to x_0, indicated by $\mathcal{N}_0$, and estimates the conditional probability for each class j as the fraction of points in $\mathcal{N}_0$ whose response values equal j

$$\Pr(Y = j \mid X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} \mathbb{1}(y_i = j). \tag{7}$$

It then assigns x_0 to the class j with the largest probability.

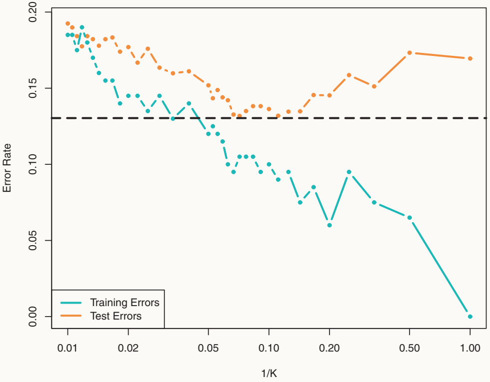
The Bias-Variance Trade-Off

KNN Applied to the Simulated Data



(Source: James et al. 2013, 41)

As $1/K$ increases, KNN becomes more flexible. A flexible KNN has low bias but high variance, while a less flexible KNN has lower variance but higher bias.



(Source: James et al. 2013, 42)

Cross-Validation Revisited

- As for regression problems, the level of flexibility is critical to the performance of a classification method.
- We can again use cross-validation to choose the optimal level of flexibility.
- However, instead of using MSE to quantify test error, we now use the number of misclassified observations.
- In the classification setting, the CV estimate for the test error rate is

$$\mathsf{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \mathsf{Err}_i, \tag{8}$$

where Err_i is the test error rate given by Equation (6).