RECSM Summer School: Machine Learning for Social Sciences

Session 1.3: Assessing Model Accuracy

Reto Wüest

Department of Political Science and International Relations
University of Geneva

Outline

1 Selection of a Machine Learning Method

2 Performance Assessment in Regression Problems
   Estimating the Performance of a Method
   The Bias-Variance Trade-Off
   Cross-Validation
   Validation Set Approach

3 Performance Assessment in Classification Problems
   Estimating the Misclassification Error of a Method
   The Bias-Variance Trade-Off
   Cross-Validation Revisited

Selection of a Machine Learning Method

Selection of a Machine Learning Method

- Our goal is to find a learning method $\hat{f}(X)$ to predict output Y on the basis of a set of inputs X.
- There are many methods available, so the question becomes how we should select $\hat{f}(X)$.
- Is there perhaps a "universal" method that performs well on all learning tasks?

Selection of a Machine Learning Method

No-Free-Lunch Theorem

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

Selection of a Machine Learning Method

- When choosing among learning methods for a given data set, we are interested in the methods' generalization performance.
- The generalization performance of a learning method relates to its prediction capability on independent test data.
- Assessment of generalization performance is very important, since it guides our choice of method for a learning task.

Performance Assessment in Regression Problems

Estimating the Performance of a Method

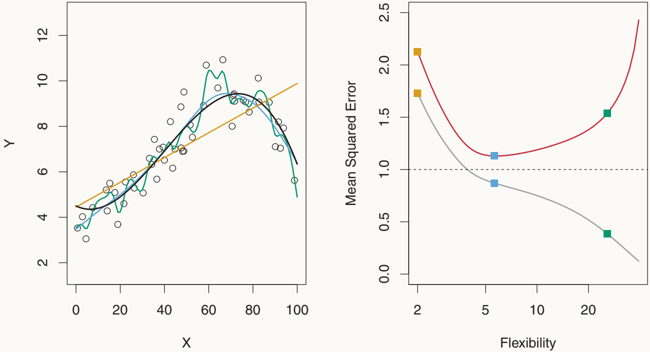
- In regression problems, the most commonly used performance measure is the mean squared error (MSE)

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{f}(x_i)\right)^2, \qquad (1.3.1)$$

  where $\hat{f}(x_i)$ is the prediction that $\hat{f}$ produces for the ith observation.
- The MSE in (1.3.1) 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 $\{(\widetilde{x}_i, \widetilde{y}_i)\}_{i=1,\ldots,M}$.
- We therefore select the method that minimizes the expected test MSE

$$\text{expected test MSE} = \frac{1}{M}\sum_{i=1}^{M}\left(\widetilde{y}_i - \hat{f}(\widetilde{x}_i)\right)^2. \qquad (1.3.2)$$

- What happens if we would select the method that minimizes the training MSE in (1.3.1)?
- Danger of overfitting data: a model that is less flexible than the one we selected would have yielded a smaller test MSE.



(Left: data simulated from true f in black; orange, blue, and green curves are three estimates for f with increasing levels of flexibility. Right: training MSE in gray; test MSE in red. Source: James et al. 2013, 31)

The Bias-Variance Trade-Off

- The U-shape in the test MSE curve is the result of two competing properties of learning methods.
- Suppose $Y = f(X) + \varepsilon$, where $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$.
- The expected test MSE of $\hat{f}(X)$ at $X = \widetilde{x}$ can be decomposed into the sum of three quantities

$$
\begin{aligned}
\text{expected test MSE} &= E\left[(Y - \hat{f}(\widetilde{x}))^2 \,\middle|\, X = \widetilde{x}\right] \qquad (1.3.3)\\
&= \left[E\left(\hat{f}(\widetilde{x})\right) - f(\widetilde{x})\right]^2 \\
&\quad + E\left[\hat{f}(\widetilde{x}) - E\left(\hat{f}(\widetilde{x})\right)\right]^2 + \sigma^2 \\
&= \text{Bias}^2\left(\hat{f}(\widetilde{x})\right) + Var\left(\hat{f}(\widetilde{x})\right) + \sigma^2,
\end{aligned}
$$

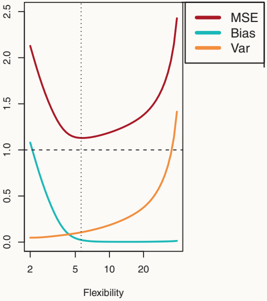
where σ^2 is the variance of the target around its true mean $f(\widetilde{x})$ (irreducible error).

The Bias-Variance Trade-Off

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

The Bias-Variance Trade-Off

- 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)

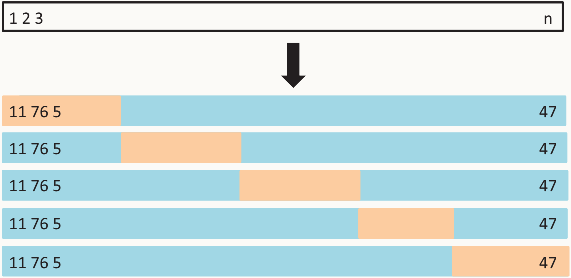
- In practice f is unobserved, making it impossible to explicitly compute the test MSE, bias, or variance for a method.
- We need to estimate the expected test MSE based on the available data (cross-validation, validation set approach).

Cross-Validation

- Cross-validation (CV) is a re-sampling method that can be used to estimate the expected test error of a learning method.
- 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 held out 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, $\text{MSE}_1, \text{MSE}_2, \ldots, \text{MSE}_K$.



(Source: James et al. 2013, 181)

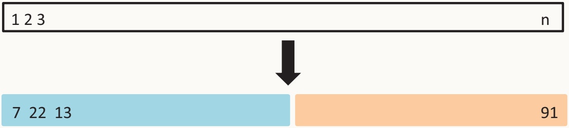
- The CV estimate for the test MSE is given by the average

$$\text{CV}_{(K)} = \frac{1}{K}\sum_{k=1}^{K} \text{MSE}_k. \tag{1.3.4}$$

- 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).
- Choice of K is associated with a bias-variance trade-off: LOOCV has lower bias than K-fold CV, but K-fold CV has lower variance than LOOCV.

Validation Set Approach

- In a data-rich situation, we can use the validation set approach to estimate the test error.
- Randomly split the N available observations into two groups, a training set and a validation set.
- Fit the model on the observations in the training set.
- Use the fitted model to predict the outcomes for the observations in the validation set and compute the MSE.



(Source: James et al. 2013, 181)

Performance Assessment in Classification Problems

Estimating the Misclassification Error of a Method

- Suppose that we estimate f on the basis of training data $\{(x_i, y_i)\}_{i=1,\ldots,n}$, where $y_1, \ldots, y_n$ are qualitative.
- The most common approach for measuring the performance of $\hat{f}$ is the misclassification error

$$\text{misclassification error} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(y_i \neq \widehat{y}_i), \qquad (1.3.5)$$

  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 misclassification error in (1.3.5) is the training error because it is computed based on the training data.

Estimating the Misclassification Error of a Method

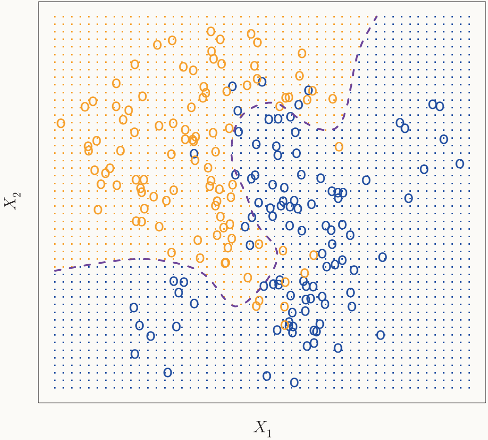
- Again, however, we are more interested in selecting a method that minimizes the expected test error

$$\text{expected test error} = \frac{1}{M}\sum_{i=1}^{M} \mathbb{1}(\widetilde{y}_i \neq \widehat{\widetilde{y}}_i). \qquad (1.3.6)$$

- One can show that the expected test error 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 expected test error (called the Bayes error rate).
- The Bayes error rate is analogous to the irreducible error in the regression setting.

Estimating the Misclassification Error of a Method

Bayes Classifier on Simulated Data



(For each $X = x$, there is a probability that Y is orange or blue. Because the data-generating process is known, the conditional probability of each class can be calculated for each x. 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. Source: James et al. 2013, 38.)

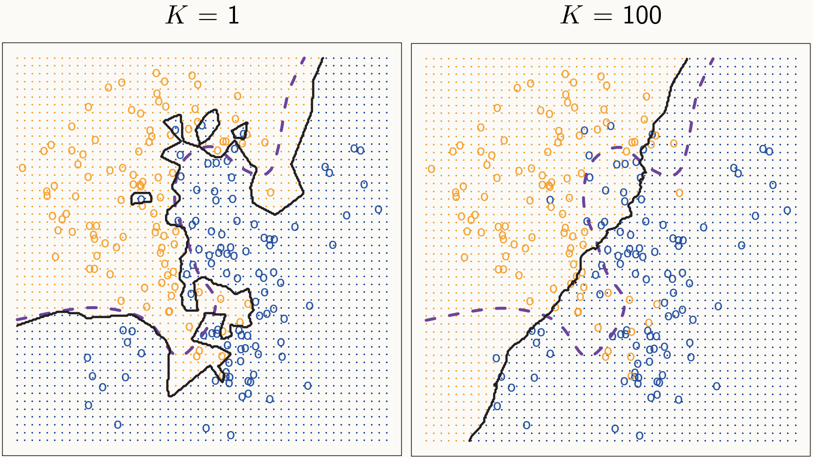
Estimating the Misclassification Error of a Method

- 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 $\widetilde{x}$, KNN identifies the K points in the training data closest to $\widetilde{x}$, indicated by $\mathcal{N}_K(\widetilde{x})$, and estimates the conditional probability for each class j as the fraction of points in $\mathcal{N}_K(\widetilde{x})$ whose response values equal j

$$\widehat{\Pr}(Y = j \mid X = \widetilde{x}) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_K(\tilde{x})} \mathbb{1}(y_i = j). \quad (1.3.7)$$

It then assigns $\widetilde{x}$ to the class j with the largest probability.

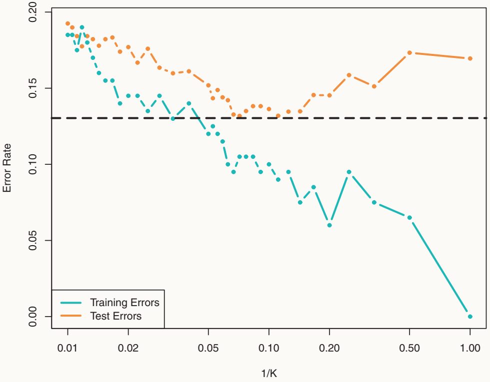
KNN Applied to the Simulated Data



(KNN decision boundaries are shown as black solid lines; Bayes decision boundary is shown as a dashed line. Source: James et al. 2013, 41)

The Bias-Variance Trade-Off

As $1/K$ increases, KNN becomes more flexible. As flexibility increases, the training error consistently declines and the test error exhibits the characteristic U-shape.



(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 expected test error is

$$\mathsf{CV}_{(K)} = \frac{1}{K}\sum_{k=1}^{K} \mathsf{Err}_k, \qquad (1.3.8)$$

where $\mathsf{Err}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} \mathbb{1}(y_i \neq \hat{y}_i)$ and N_k is the number of observations in the kth validation set.