Chapter 5 – Cross validation

Wenjing Liao

School of Mathematics
Georgia Institute of Technology

Math 4803
Fall 2019

Outline

1. Review – the bias and variance trade off
2. 5.1.1 Validation set
3. 5.1.2 Leave-One-Out Cross-Validation
4. 5.1.5 Cross-Validation and Classification Problems

The bias-variance trade off

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

Variance is the amount by which $\hat{f}$ would change if we estimated it using a different training data set.

Bias refers to the error that is introduced by approximating the complicated ground-truth f by a simpler model.

KNN classification

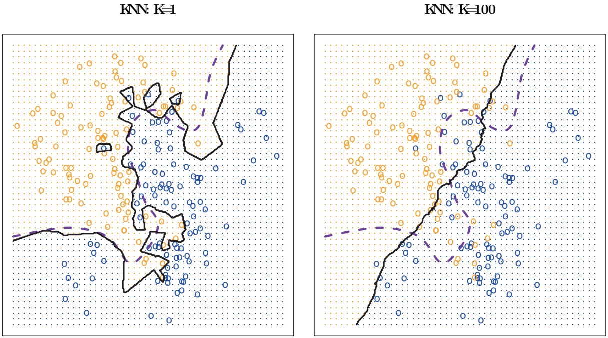

FIGURE 2.16. *A comparison of the KNN decision boundaries (solid black curves) obtained using* $K = 1$ *and* $K = 100$ *on the data from Figure 2.13. With* $K = 1$*, the decision boundary is overly flexible, while with* $K = 100$ *it is not sufficiently flexible. The Bayes decision boundary is shown as a purple dashed line.*

- $K = 1$: a small bias and a large variance
- $K = 100$: a large bias and a small variance

Training and test error versus K

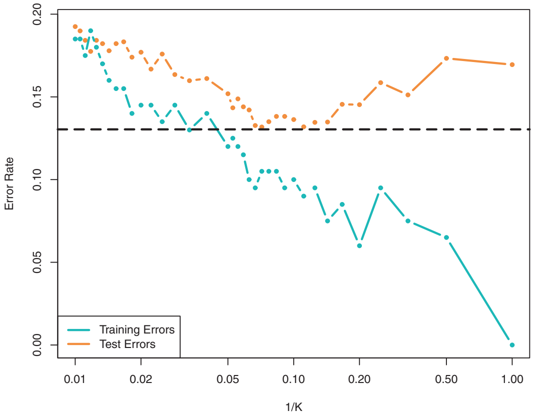

FIGURE 2.17. *The KNN training error rate (blue, 200 observations) and test error rate (orange, 5,000 observations) on the data from Figure 2.13, as the level of flexibility (assessed using $1/K$) increases, or equivalently as the number of neighbors K decreases. The black dashed line indicates the Bayes error rate. The jumpiness of the curves is due to the small size of the training data set.*

Question: How to choose a proper K?

Fit data with polynomials

Linear: $Y = \beta_0 + \beta_1 X$

Quadratic: $Y = \beta_0 + \beta_1 X + \beta_2 X^2$

Polynomials of degree p: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_p X^p$

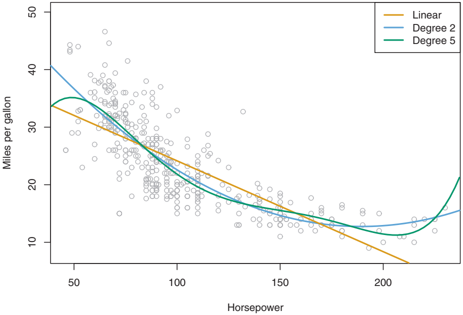

FIGURE 3.8. *The* Auto *data set. For a number of cars,* mpg *and* horsepower *are shown. The linear regression fit is shown in orange. The linear regression fit for a model that includes* horsepower^2 *is shown as a blue curve. The linear regression fit for a model that includes all polynomials of* horsepower *up to fifth-degree is shown in green.*

Question: How to choose a proper p?

Outline

1. Review – the bias and variance trade off
2. 5.1.1 Validation set
3. 5.1.2 Leave-One-Out Cross-Validation
4. 5.1.5 Cross-Validation and Classification Problems

Training and validation set

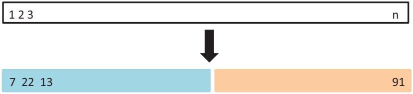

FIGURE 5.1. *A schematic display of the validation set approach. A set of n observations are randomly split into a training set (shown in blue, containing observations 7, 22, and 13, among others) and a validation set (shown in beige, and containing observation 91, among others). The statistical learning method is fit on the training set, and its performance is evaluated on the validation set.*

Polynomial order in mpg ∼ horsepower:

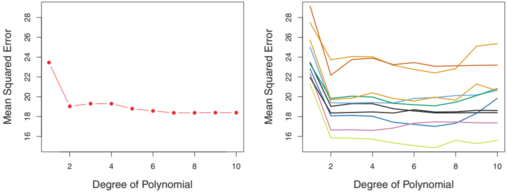

FIGURE 5.2. *The validation set approach was used on the* `Auto` *data set in order to estimate the test error that results from predicting* `mpg` *using polynomial functions of* `horsepower`. Left: *Validation error estimates for a single split into training and validation data sets.* Right: *The validation method was repeated ten times, each time using a different random split of the observations into a training set and a validation set. This illustrates the variability in the estimated test MSE that results from this approach.*

Outline

1. Review – the bias and variance trade off
2. 5.1.1 Validation set
3. 5.1.2 Leave-One-Out Cross-Validation
4. 5.1.5 Cross-Validation and Classification Problems

Leave-One-Out Cross-Validation (LOOCV)

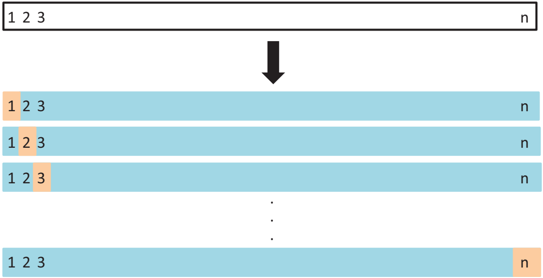

FIGURE 5.3. *A schematic display of LOOCV. A set of n data points is repeatedly split into a training set (shown in blue) containing all but one observation, and a validation set that contains only that observation (shown in beige). The test error is then estimated by averaging the n resulting MSE's. The first training set contains all but observation 1, the second training set contains all but observation 2, and so forth.*

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{MSE}_i.$$

LOOCV

Expensive to implement: one needs to fit the model n times.

Least squares linear or polynomial regression:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2,$$

where $\hat{y}_i$ is the ith fitted value from the original least squares fit, and h_i is a leverage quantity (see Page 98):

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}.$$

The leverage h_i lies between $1/n$ and 1, and reflects the amount that an observation influences its own fit.

About leverage and outlier

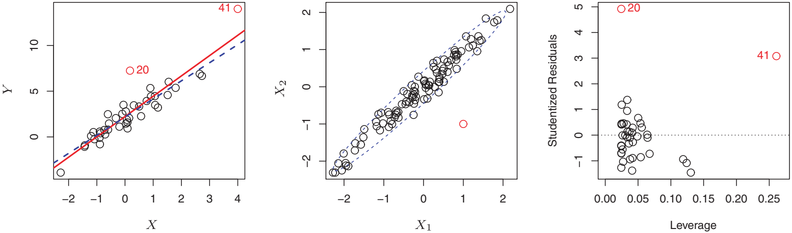

FIGURE 3.13. Left: *Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 41 removed.* Center: *The red observation is not unusual in terms of its X_1 value or its X_2 value, but still falls outside the bulk of the data, and hence has high leverage.* Right: *Observation* 41 *has a high leverage and a high residual.*

k-Fold Cross-Validation

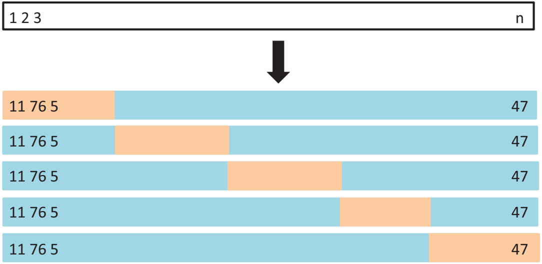

FIGURE 5.5. *A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.*

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

It is common to set $k = 5$ or $k = 10$.

Degree of polynomial

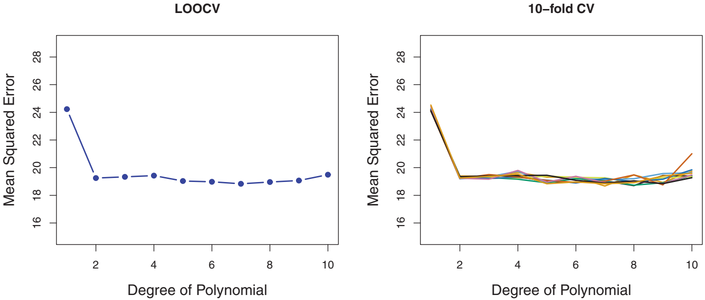

FIGURE 5.4. *Cross-validation was used on the* `Auto` *data set in order to estimate the test error that results from predicting* `mpg` *using polynomial functions of* `horsepower`. Left: *The LOOCV error curve.* Right: 10-*fold CV was run nine separate times, each with a different random split of the data into ten parts. The figure shows the nine slightly different CV error curves.*

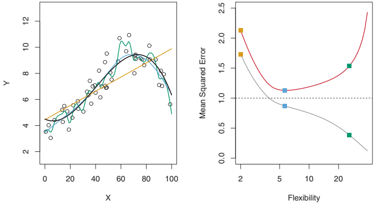

FIGURE 2.9. Left: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.*

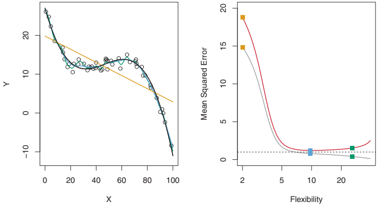

FIGURE 2.11. *Details are as in Figure 2.9, using a different f that is far from linear. In this setting, linear regression provides a very poor fit to the data.*

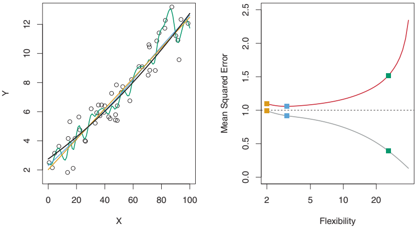

FIGURE 2.10. *Details are as in Figure 2.9, using a different true f that is much closer to linear. In this setting, linear regression provides a very good fit to the data.*

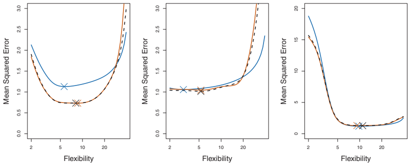

FIGURE 5.6. *True and estimated test MSE for the simulated data sets in Figures 2.9* (left)*, 2.10* (center)*, and 2.11* (right)*. The true test MSE is shown in blue, the LOOCV estimate is shown as a black dashed line, and the 10-fold CV estimate is shown in orange. The crosses indicate the minimum of each of the MSE curves.*

Outline

1. Review – the bias and variance trade off
2. 5.1.1 Validation set
3. 5.1.2 Leave-One-Out Cross-Validation
4. 5.1.5 Cross-Validation and Classification Problems

CV in the classification setting

LOOCV:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Err}_i,$$

where $\mathrm{Err}_i = I(y_i \neq \hat{y}_i)$

Example with logistic regression:

Linear logistic regression

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Quadratic logistic regression

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2$$

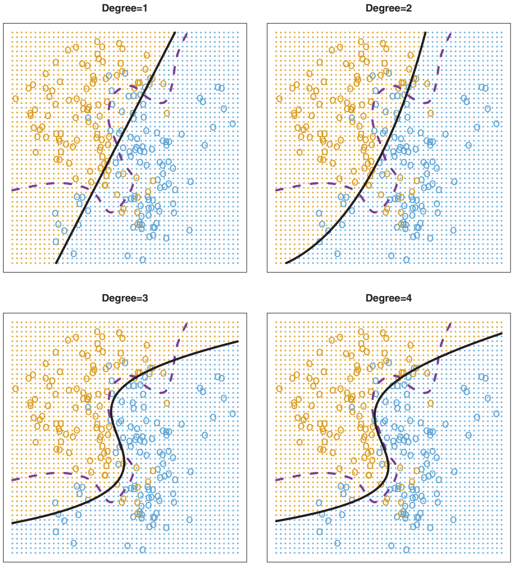

FIGURE 5.7. *Logistic regression fits on the two-dimensional classification data displayed in Figure 2.13. The Bayes decision boundary is represented using a purple dashed line. Estimated decision boundaries from linear, quadratic, cubic and quartic (degrees 1–4) logistic regressions are displayed in black. The test error rates for the four logistic regression fits are respectively* 0.201, 0.197, 0.160, *and* 0.162, *while the Bayes error rate is* 0.133.

Cross-Valication

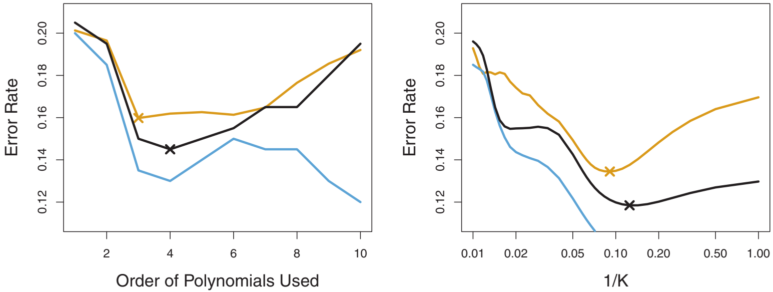

FIGURE 5.8. *Test error (brown), training error (blue), and* 10*-fold CV error (black) on the two-dimensional classification data displayed in Figure 5.7.* Left: *Logistic regression using polynomial functions of the predictors. The order of the polynomials used is displayed on the x-axis.* Right: *The KNN classifier with different values of K, the number of neighbors used in the KNN classifier.*

Reference

Textbook: James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani, An introduction to statistical learning. Vol. 112, New York: Springer, 2013