IEOR165 Discussion
Week 6

Sheng Liu

University of California, Berkeley

Feb 26, 2016

Outline

Corss-Validation: Motivation

Model assessment: evaluate the performance of a model

- True error: the error rate on the entire population
- Training error: the error rate on the training data set
- Test error: the average error that results from using a statistical learning method to predict the response on a new observation—that is, a measurement that was not used in training the method

Suppose we have observations $(x_1, y_1), \ldots, (x_n, y_n)$, we have constructed a linear regression model based on these observations (training set) : $\hat{y} = \hat{b} + \hat{m}x$. Now for a random new data point (x, y), the test error is

$$\mathbb{E}[(y - \hat{b} - \hat{m}x)^2]$$

Corss-Validation: Motivation

Model selection

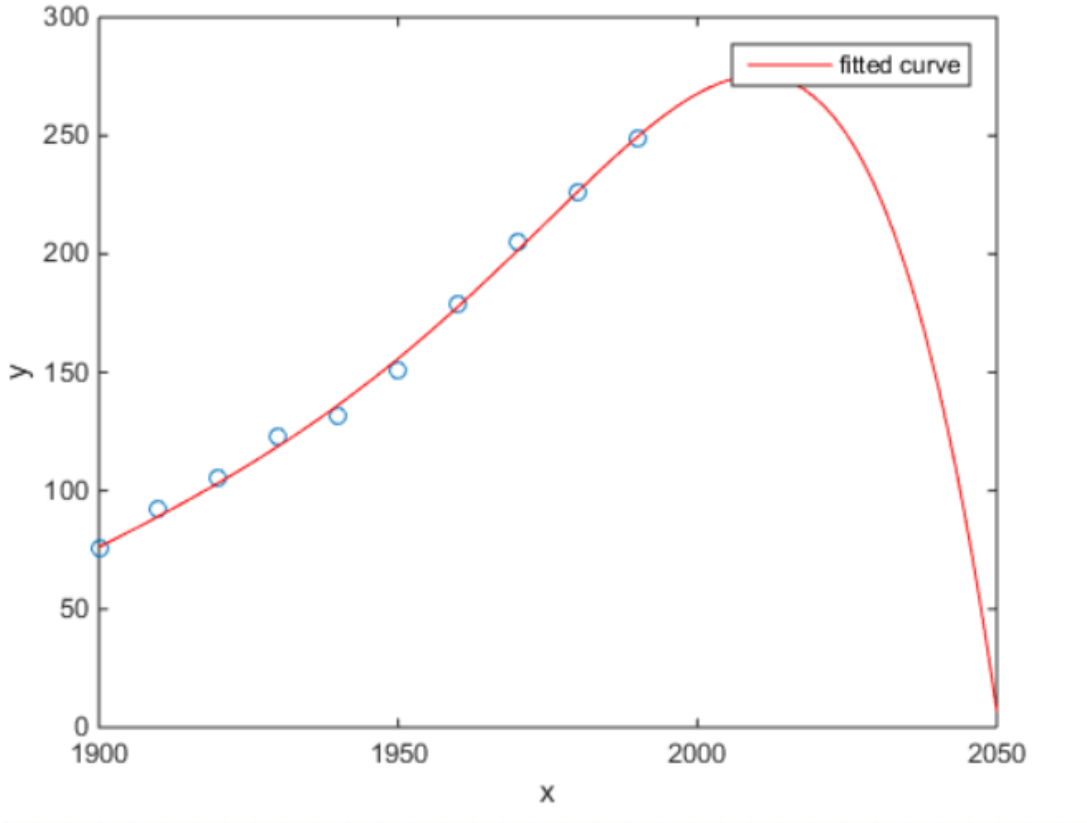
- Select the proper (tuning) parameters for a model: λ in the ridge regression
- Select linear regression versus polynomial regression
- Compare test errors

Why compare test errors?

- We do not know the true error
- Training error is too optimistic
  - When fitting a model to the training data, we are minimizing the training error, it is quite often to see very high accuracy on training data set
  - The model may overfit the training data, which causes poor predictability

Corss-Validation: Motivation

Overfitting: US population versus time



Source: `http://www.mathworks.com/help/curvefit/examples/polynomial-curve-fitting.html?requestedDomain=www.mathworks.com#zmw57dd0e115`

Validation set

How to estimate the test error? A naive way is to split the data set into two parts:

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.*

Drawback: the test error is depending on which observations are included in the training set and thus is highly variable, we may get a poor result just because of a poor split

Leave-One-Out Cross-Validation

Berkeley UNIVERSITY OF CALIFORNIA

We can repeat the split many times. Each time, we take one data point as the validation set.

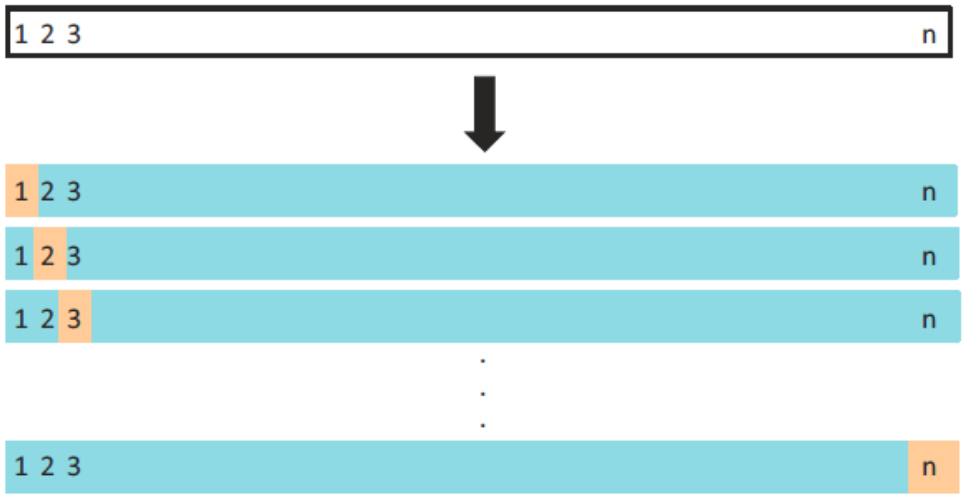


FIGURE 5.3. *A schematic display of LOOCV. A set of n data points is repeatedly split into a training set (shown in blue) containing all but one observation, and a validation set that contains only that observation (shown in beige). The test error is then estimated by averaging the n resulting MSE's.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

Leave-One-Out Cross-Validation

The test error is estimated by

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

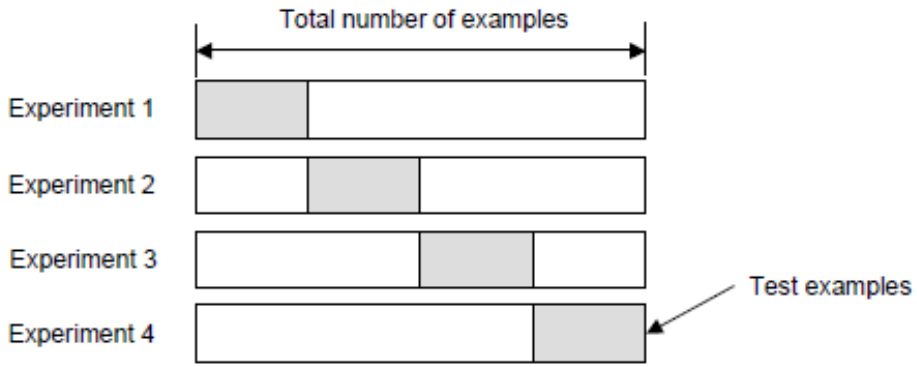
where $\hat{y}_i$ is estimated by the model using the training data set

$$(x_1, y_1), \ldots, (x_{i-1}, y_{i-1}),\ (x_{i+1}, y_{i+1}), \ldots, (x_n, y_n)$$

Can be generalized to Leave-k-Out Cross-Validation

k-Fold Cross-Validation

- Randomly divide the data set into k groups, or folds, of approximately equal size
- Let the ith fold as the validation set and the remaining (k-1) folds as training set. Fit the model to the training data set and use it to make predictions on the validation set. Then the mean squared error can be calculated on the validation set
- Repeat this process k times and select a different fold as the validation set each time



Comparisons

- Leave-One-Out Cross-Validation is a special case of k-Fold Cross-Validation
- Why choose $k < n$?
  - Computational feasibility: fitting k times or n times
  - Bias-Variance trade-off: smaller k increases the bias but decrease the variance
- 5-folds and 10-folds are quite popular

CDF Estimation

- Parametric statistics: assume the data follows certain distribution which we can characterize by a number of parameters. We will first estimate the parameters and then gives the corresponding CDF.
- Nonparametric statistics: try to make no assumptions about the distribution of the data, i.e. CDF can take any forms as long as it meets the definition of CDF. We will estimate the CDF directly from the data.
- Empirical distribution function

$$F(u) = P(X \leq u)$$

$$\hat{F}(u) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq u)$$

PDF Estimation

- Empirical distribution function is not continuous and thus not differentiable, its pdf can be defined using Dirac delta function
- Use histograms
  - Specify the support of the distribution (range) and divide it into several bins
  - Count the number of data points in each bin

$$\hat{f}(u) = \frac{1}{n} \frac{\#\text{ of data points in the same bin as } u}{\text{width of bin}}$$

  - Drawbacks: discontinuity, curse of dimensionality

PDF Estimation

- Another naive estimator:

$$\hat{f}(u) = \frac{1}{n}\frac{\#\text{ of data points in } N(u)}{\text{width of } N(u)}$$
$$= \frac{1}{n \cdot \text{width of } N(u)} \sum_{i=1}^{n} I(x_i \in N(u))$$

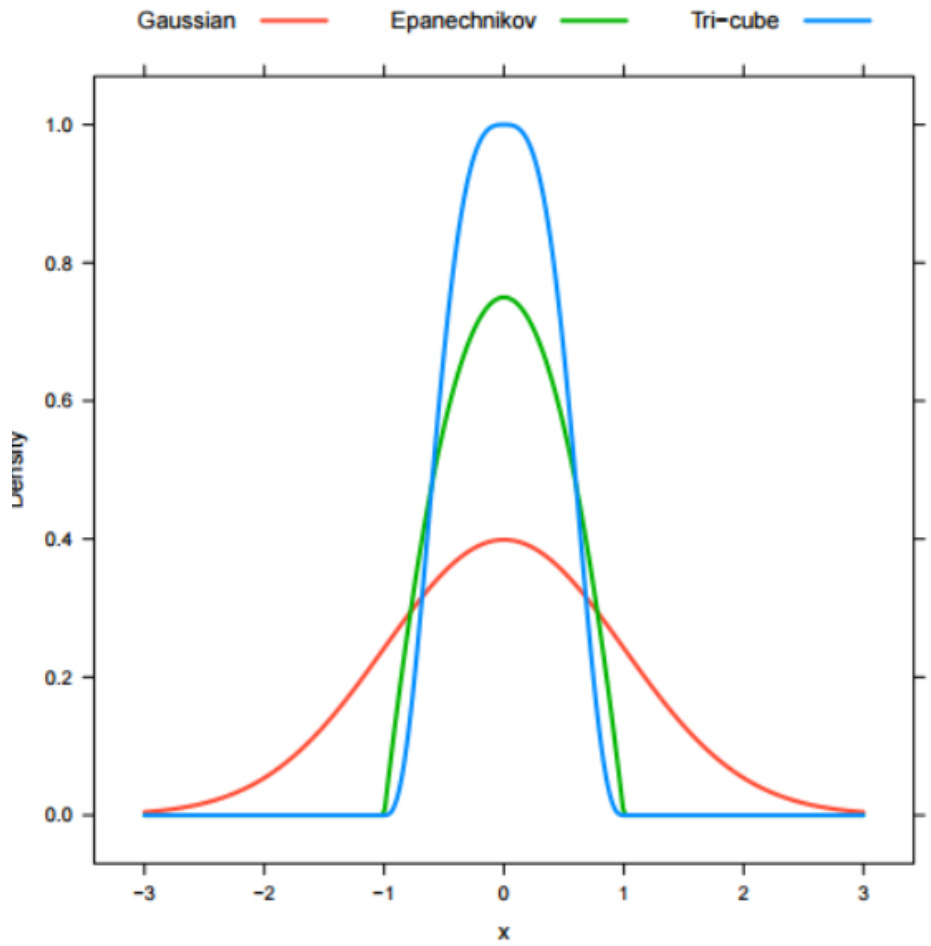
  where $N(u)$ is a small metric neighborhood of u, it can be a hypercube in the 3-dimensional space. Again, this is not continuous.

- Kernel density estimate:

$$\hat{f}(u) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{u - x_i}{h})$$

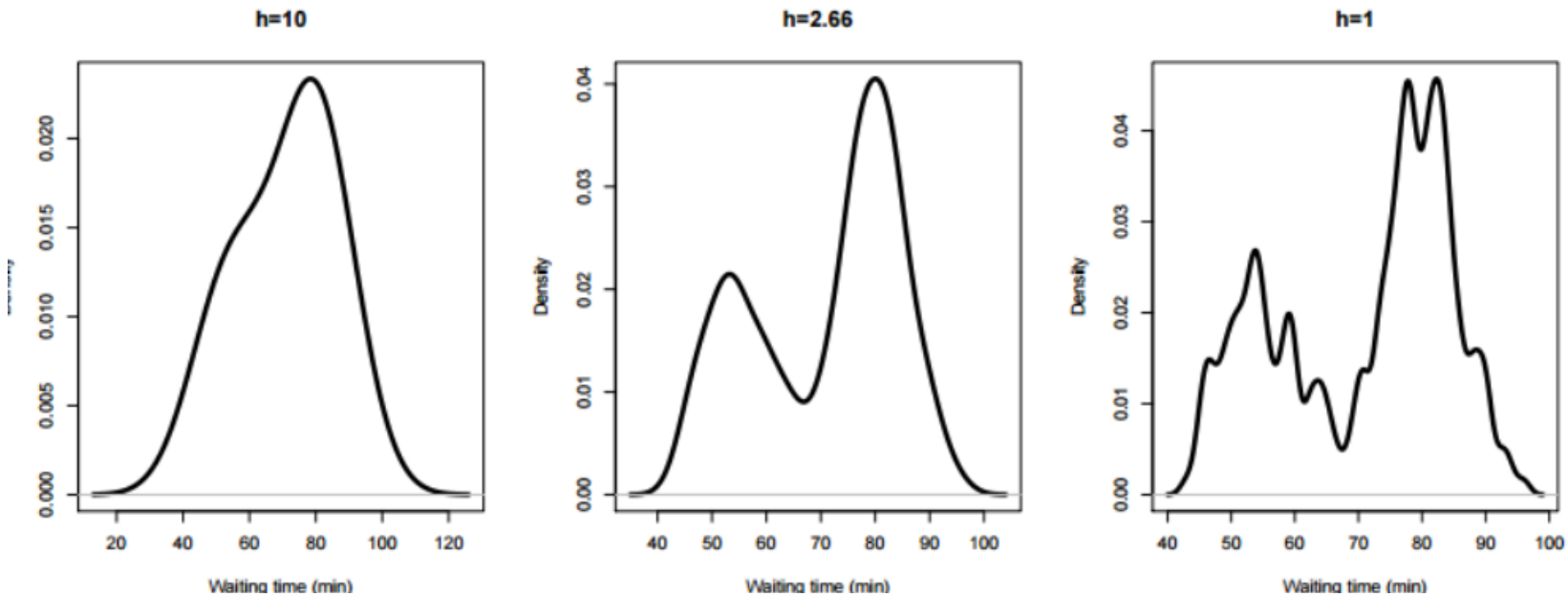
  where we call h as bandwidth.

Kernel functions



Kernel functions

- Kernel function provides smoothness
- Selection of bandwidth controls the degree of smoothing



References

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). New York: springer.
Ricardo Gutierrez-Osuna:
http://courses.cs.tamu.edu/rgutier/csce666_f13/
http://research.cs.tamu.edu/prism/lectures/iss/iss_l13.pdf