STAT 339
Cross-Validation

February 7, 2020

Colin Reimer Dawson

Questions/Administrative Business

- HW1 will be posted this weekend (I'll send a Slack announcement when it's up)

Outline

Evaluating a Classifier
- Validation and Test Sets
- K-fold Cross-Validation

Evaluating a Supervised Learning Method

Two Kinds of Evaluation

1. How do we decide which free "parameters", like the K in KNN, are best?
2. How do we know how good a job our final method has done?

Two Choices To Be Made

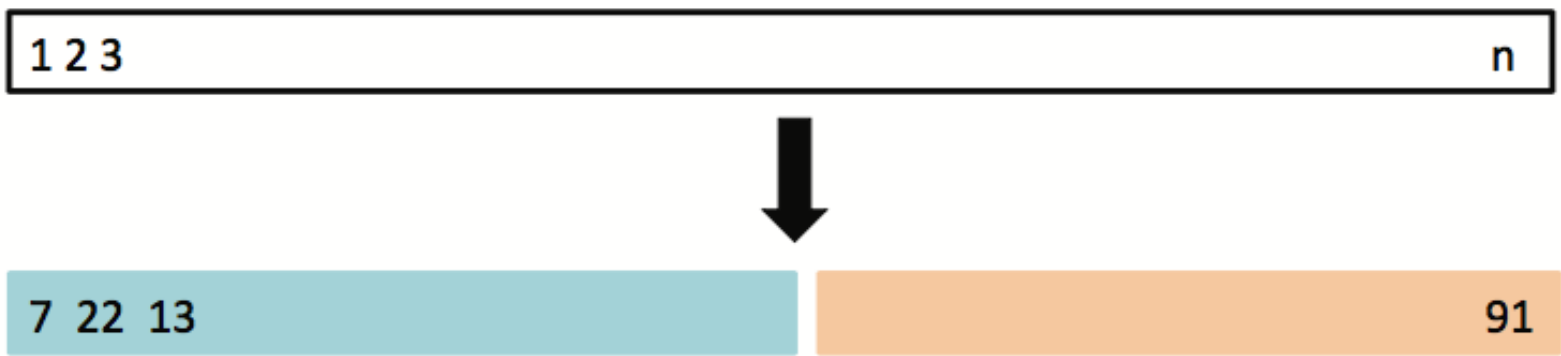
1. **How do we quantify** performance?
2. **What data** do we use to measure performance?

Overfitting and Test Set

- Fitting and evaluating on the **same data** usually results in **overfitting**.
- **Overfitting** is mistaking noise for signal, and trying to learn patterns in noise that are illusions
- To avoid overfitting, use **different data** for evaluation vs. fitting. This "held out data" is called a **validation/test set**

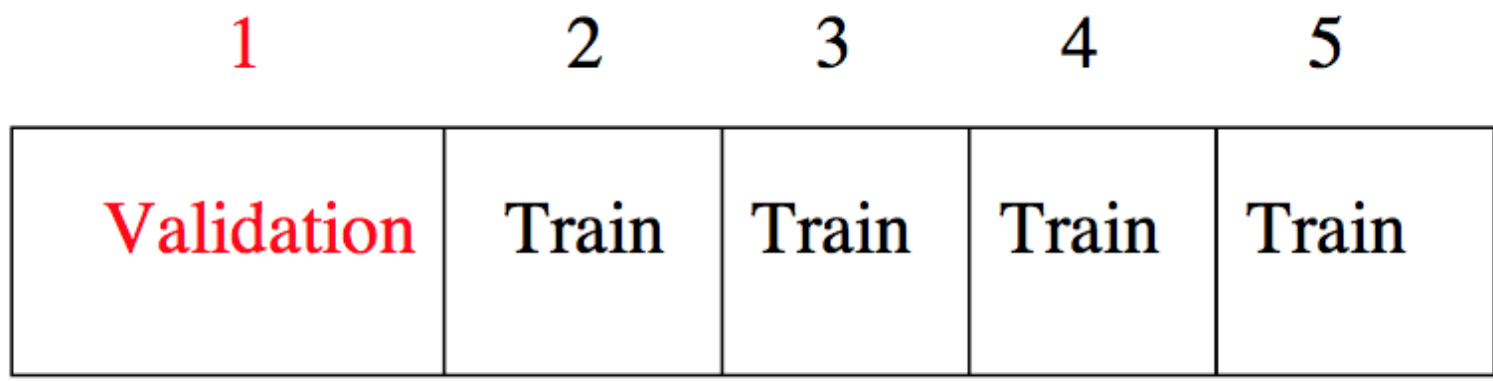
Validation vs. Test Set

- If we select the best version of our method by **optimizing performance on the test set, we have no objective measure of absolute performance**
- Performance of the **best model** on the test set is **overly optimistic**
- Instead, randomly subdivide the **training set** into **training** and **validation** sets.
- Use **training to do classification**; **validation to evaluate** and guide "higher-order" decisions.



K-fold Cross-Validation

- Sacrificing training data: noisy learning
- Sacrificing validation data: noisy evaluation
- K-fold Cross-Validation
  - Divide training set into K equal parts, or "folds"; each fold serves as validation set
  - Report generalization error averaged over folds



- $K = N$: "Leave-one-out" Cross-validation

K-fold Cross Validation Algorithm

A. For each method, $\mathcal{M}_j$, under consideration ($j = 1, \ldots, J$)

1. Divide training set into K "folds" with (approximately) equal cases per fold. (Keep test set "sealed")
2. For $k = 1, \ldots, K$:
   (a) Designate fold k the "validation set", the rest are the training set
   (b) "Train" the algorithm on the training set, obtaining classification function c_k
   (c) compute error rate, Err_k on the validation set

$$\mathrm{Err}_k(\mathcal{M}_j) = \frac{1}{|\text{Validation}|} \sum_{n \in \text{Validation}} I(c_k(\mathbf{x}_n) \neq t_n)$$

3. Return mean error rate across folds

$$\overline{\mathrm{Err}}(\mathcal{M}_j) = \frac{1}{K} \sum_{k=1}^{K} Err_k(\mathcal{M}_j)$$

B. Select $\mathcal{M}_j$ with lowest $\overline{\mathrm{Err}}$: $j = \arg\min \overline{\mathrm{Err}}(\mathcal{M}_j)$