PS5841

Data Science in Finance & Insurance

KNN

Yubo Wang

Autumn 2022

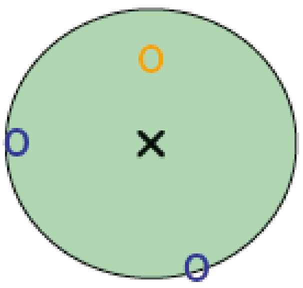
Columbia University in the City of New York

Data

- Population/Truth
- Training Set
- Validation Set
- Test Set

- Features
- Response

K Nearest Neighbors

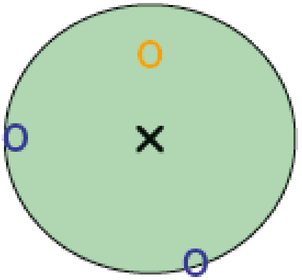


- Standardize features when appropriate

KNN Regressor

- k-nearest neighbor fit

$$\hat{Y}(\boldsymbol{x}_0) = \frac{1}{k} \sum_{\boldsymbol{x} \in N_k(\boldsymbol{x}_0)} y_i$$



Prediction: mean response in the neighborhood

- MSE

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

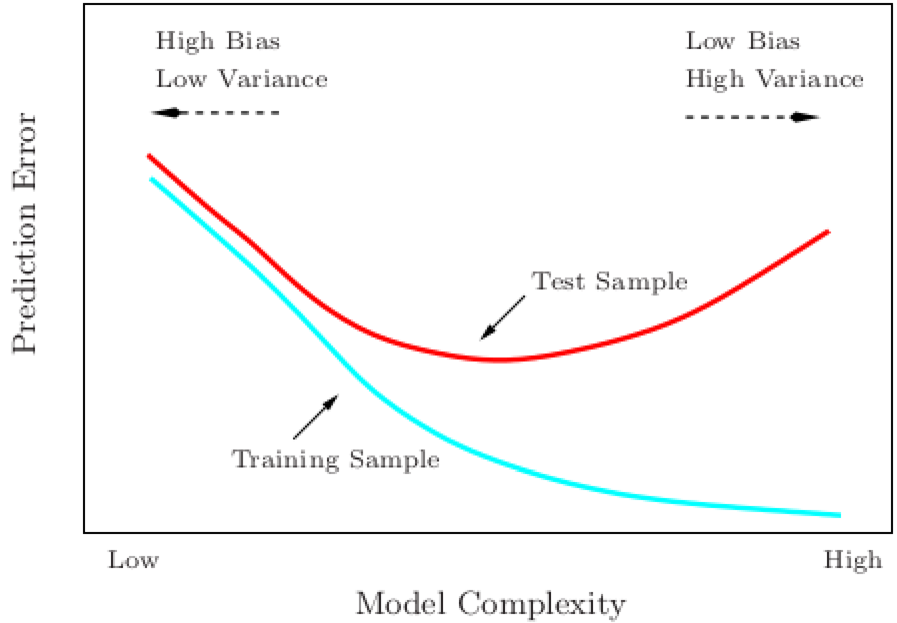
Prediction Error (regression)

- Mean Squared Error

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

- Training Error
- Test (Generalization) Error
  – Expected overall test MSE: the average test MSE resulting from fitting the model with a large number of training sets and test each on the test set.

U-Shaped Test Error



- Bias
  - How good is the model prediction on average
- Variance
  - Fluctuation in predictions resulting from fitting the model with different training sets

Bias-Variance Tradeoff (regression)

- Suppose

$$Y = f(X) + \epsilon,$$
$$E[\epsilon] = 0, \qquad Var(\epsilon) = \sigma^2$$

- Bias-Variance decomposition

$$E_{tr}[(Y - \hat{f}(x_0)^2 | X = x_0]$$

$$= E_{tr}\left\{\left[\hat{f}(x_0) - E_{tr}\left(\hat{f}(x_0)\right)\right]^2\right\}$$ variance

$$+\left[E_{tr}\left(\hat{f}(x_0)\right) - f(x_0)\right]^2$$ bias

$$+\sigma^2$$ irreducible error

Bias-Variance Tradeoff (KNN regression)

- Variance

$$\frac{\sigma^2}{k}$$

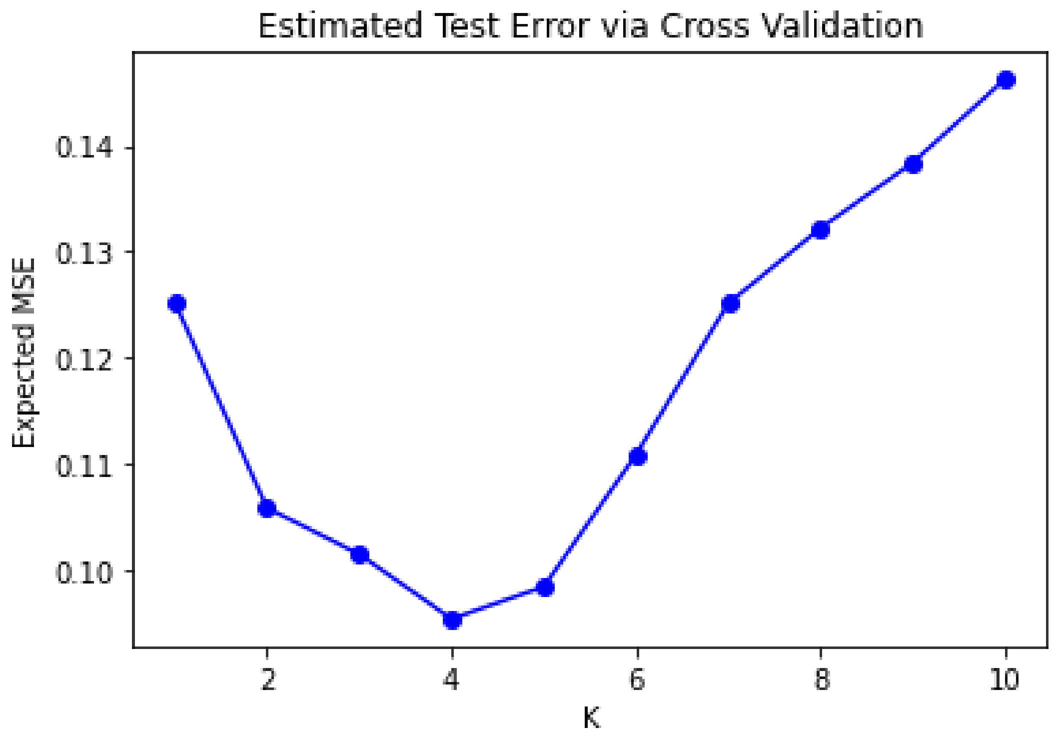
- bias²

$$+\left[\left(\frac{1}{k}\sum_{x\in N_k(\boldsymbol{x}_0)} y_i\right) - f(\boldsymbol{x}_0)\right]^2$$

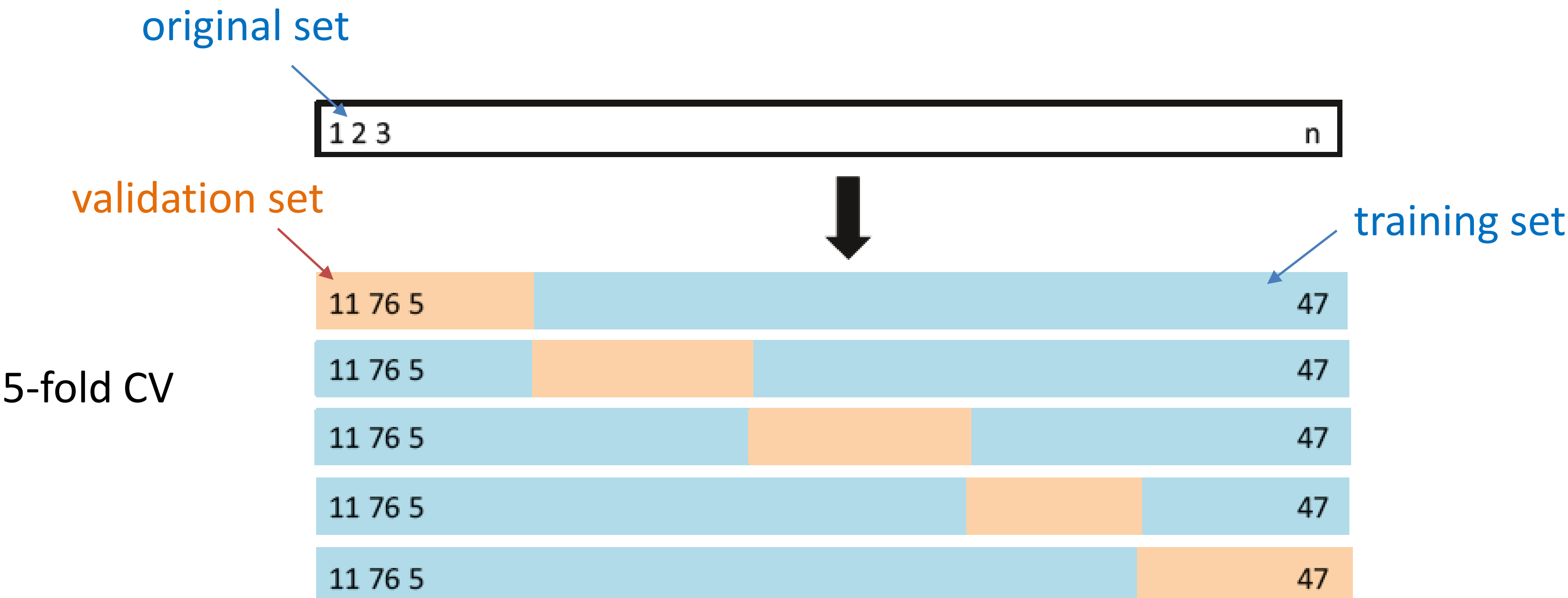
- Irreducible error

$$\sigma^2$$

Find Optimal K via Cross Validation



k-Fold Cross Validation

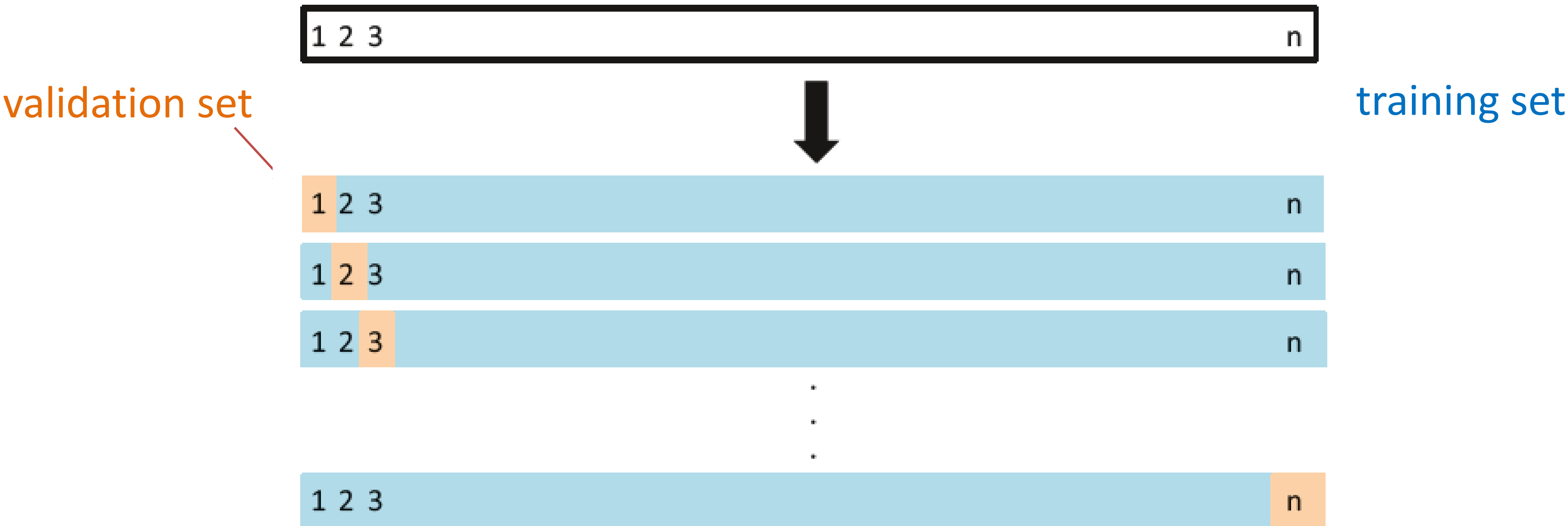


k-fold CV: For each validation set

- Fit model on $k-1$ folds (training set)
- Compute "$Error_i$" on the hold-out fold (validation set)
- Compute the CV estimate of the "test error"

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} Error_i$$

Leave-One-Out Cross Validation (LOOCV)



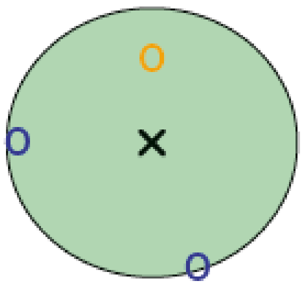
Magic Numbers

- There is a bias-variance trade-off associated with the choice of k in k-fold CV.
  - LOOCV produces less bias than k-fold CV, but with higher variance
- Empirically, k=5, or k=10
  - Neither excessively high bias nor high variance

KNN Classifier

- k-nearest neighbor fit

$$\Pr(Y = j|X = \boldsymbol{x}_0) = \frac{1}{k} \sum_{x \in N_k(\boldsymbol{x}_0)} I(y_i = j)$$

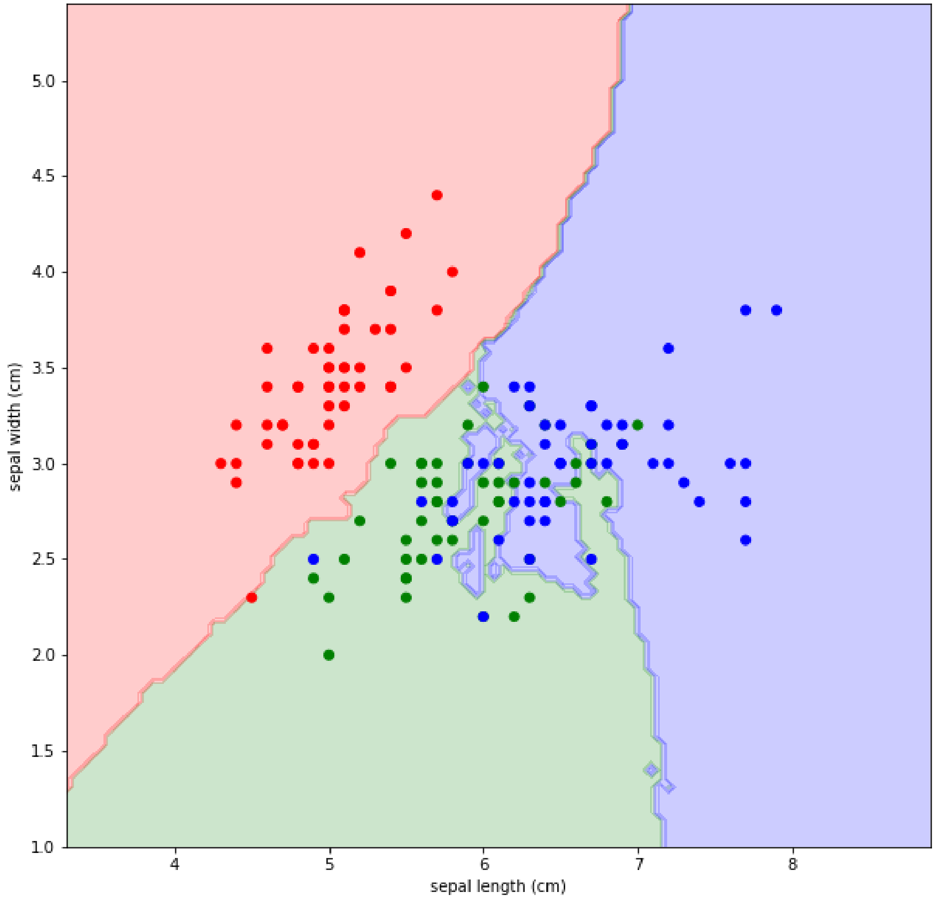


Classify $\boldsymbol{x}_0$ to the class with the highest probability

- Error rate = 1 - accuracy

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

Decision Boundary



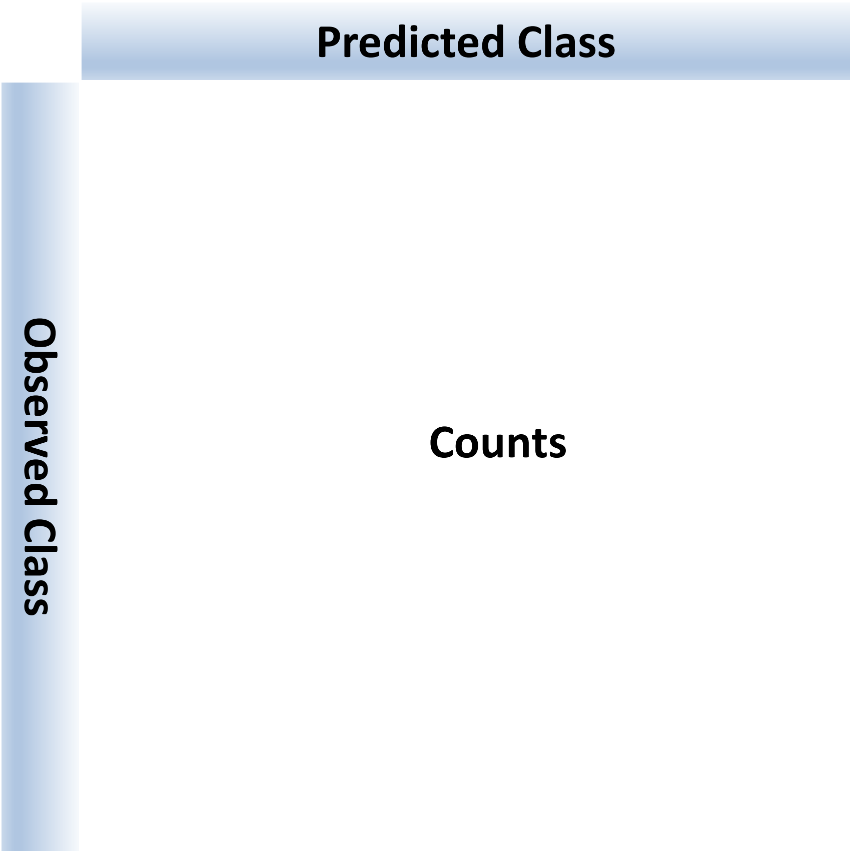
Confusion Matrix

- Accuracy

% captured by diagonal

- Error Rate

1 - Accuracy



That was