MONASH University

ETC3250

Business Analytics

Week 2.
Assessing model accuracy

3 August 2015

Outline

Assessing model accuracy

Suppose we have a regression model $y = f(x) + \varepsilon$.

Estimate $\hat{f}$ from some **training data**, $Tr = \{x_i, y_i\}_1^n$.

One common measure of accuracy is:

Training Mean Squared Error

$$\text{MSE}_{Tr} = \underset{i \in Tr}{\text{Ave}}[y_i - \hat{f}(x_i)]^2 = \frac{1}{n}\sum_{i=1}^{n}[(y_i - \hat{f}(x_i)]^2$$

Assessing model accuracy

Suppose we have a regression model $y = f(x) + \varepsilon$.

Estimate $\hat{f}$ from some **training data**, $Tr = \{x_i, y_i\}_1^n$.

One common measure of accuracy is:

Training Mean Squared Error

$$\text{MSE}_{Tr} = \underset{i \in Tr}{\text{Ave}}[y_i - \hat{f}(x_i)]^2 = \frac{1}{n}\sum_{i=1}^{n}[(y_i - \hat{f}(x_i)]^2$$

Measure real accuracy using **test data** $Te = \{x_j, y_j\}_1^m$

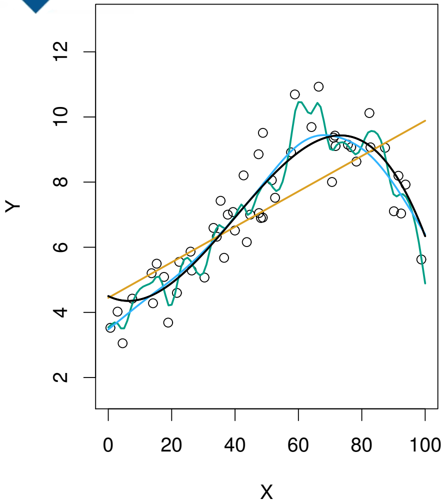
Test Mean Squared Error

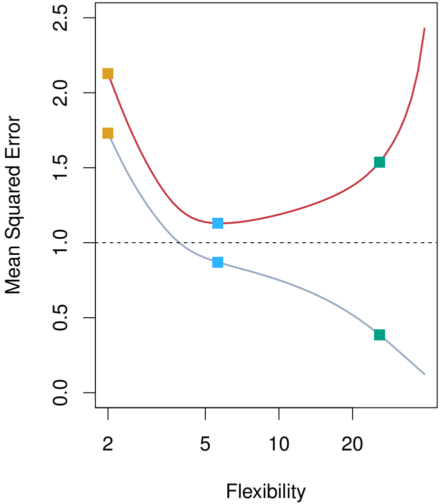
$$\text{MSE}_{Te} = \underset{j \in Te}{\text{Ave}}[y_j - \hat{f}(x_j)]^2 = \frac{1}{m}\sum_{j=1}^{m}[(y_j - \hat{f}(x_j)]^2$$

Training vs Test MSEs

- In general, the more flexible a method is, the lower its training MSE will be. i.e. it will “fit” the training data very well.
- However, the test MSE may be higher for a more flexible method than for a simple approach like linear regression.
- Flexibility also makes interpretation more difficult. There is a trade-off between flexibility and model interpretability.

Example: splines





Black: true curve

Orange: linear regression

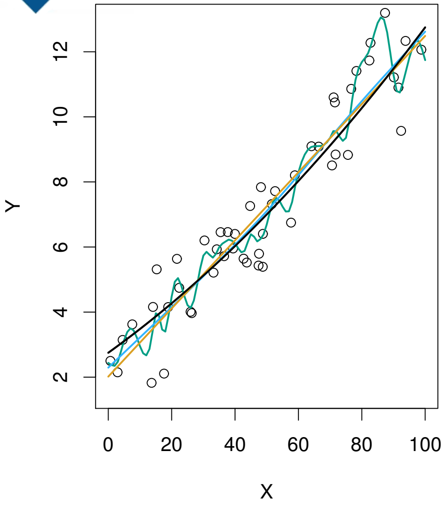
Blue/green: Smoothing splines

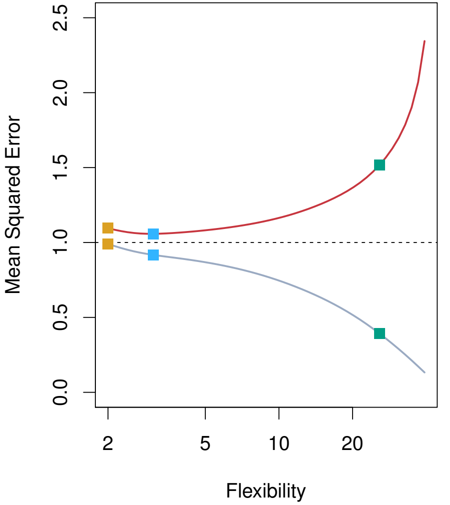
Grey: Training MSE

Red: Test MSE

Dashed: Minimum test MSE

Example: splines

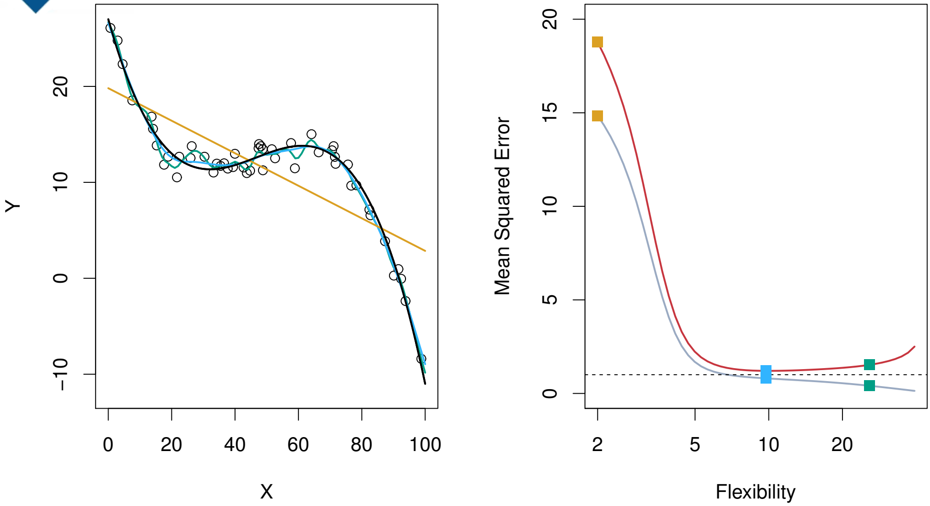




Black: true curve
Orange: linear regression
Blue/green: Smoothing splines

Grey: Training MSE
Red: Test MSE
Dashed: Minimum test MSE

Example: splines



Black: true curve
Orange: linear regression
Blue/green: Smoothing splines

Grey: Training MSE
Red: Test MSE
Dashed: Minimum test MSE

Bias-variance tradeoff

There are two competing forces that govern the choice of learning method: **bias** and **variance**.

Bias

is the error that is introduced by modeling a complicated problem by a simpler problem.

- For example, linear regression assumes a linear relationship when few real relationships are exactly linear.
- In general, the more flexible a method is, the less bias it will have.

Bias-variance tradeoff

There are two competing forces that govern the choice of learning method: **bias** and **variance**.

Variance

refers to how much your estimate would change if you had different training data.

- In general, the more flexible a method is, the more variance it has.
- The size of the training data has an impact on the variance

The bias-variance tradeoff

MSE decomposition

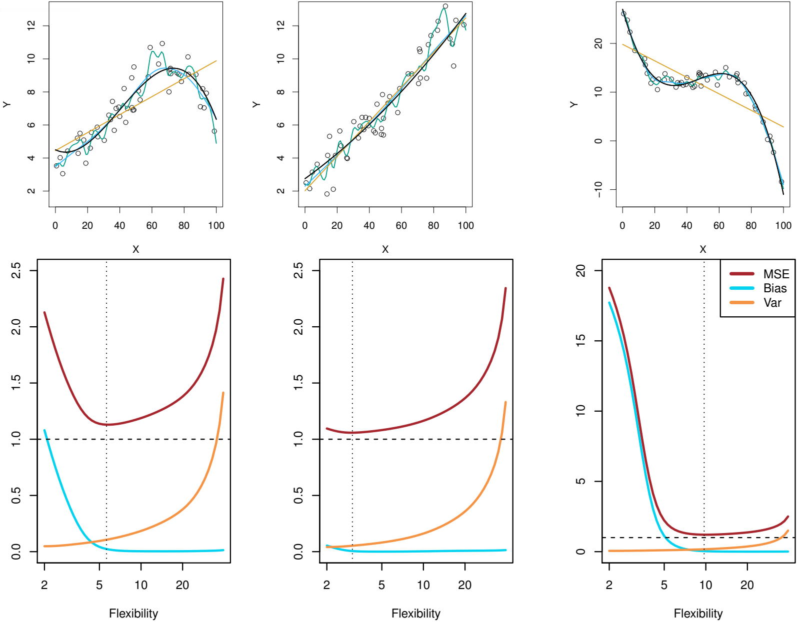
If $Y = f(x) + \varepsilon$ and $f(x) = \mathrm{E}[Y \mid X = x]$, then the expected **test** MSE for a new Y at x_0 will be equal to

$$\mathrm{E}[(Y - \hat{f}(x_0))^2] = [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\hat{f}(x_0)) + \mathrm{Var}(\varepsilon)$$

Test MSE = Bias^2 + Variance + Irreducible variance

- The expectation averages over the variability of Y as well as the variability in the training data.
- As the flexibility of $\hat{f}$ increases, its variance increases and its bias decreases.
- Choosing the flexibility based on average test MSE amounts to a **bias-variance trade-off**.

Bias-variance trade-off



Optimal prediction

MSE decomposition

If $Y = f(x) + \varepsilon$ and $f(x) = \mathrm{E}[Y \mid X = x]$, then the expected **test** MSE for a new Y at x_0 will be equal to

$$\mathrm{E}[(Y - \hat{f}(x_0))^2] = [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\hat{f}(x_0)) + \mathrm{Var}(\varepsilon)$$

The optimal MSE is obtained when

$$\hat{f} = f = \mathrm{E}[Y \mid X = x].$$

Then bias=variance=0 and

$$\mathrm{MSE} = \text{irreducible variance}$$

This is called the "oracle" predictor because it is not achievable in practice.

Outline

Classification problems

Here the response variable Y is qualitative.

- e.g., email is one of $\mathcal{C} = (\text{spam}, \text{ham})$
- e.g., voters are one of $\mathcal{C} = (\text{Liberal}, \text{Labor}, \text{Green}, \text{National}, \text{Other})$

Our goals are:

1. Build a classifier $C(x)$ that assigns a class label from $\mathcal{C}$ to a future unlabeled observation x.
2. Assess the uncertainty in each classification (i.e., the probability of misclassification).
3. Understand the roles of the different predictors among $X = (X_1, X_2, \ldots, X_p)$.

Classification problems

Here the response variable Y is qualitative.

- e.g., email is one of $\mathcal{C} = (\text{spam}, \text{ham})$
- e.g., voters are one of $\mathcal{C} = (\text{Liberal}, \text{Labor}, \text{Green}, \text{National}, \text{Other})$

Our goals are:

1. Build a classifier $C(x)$ that assigns a class label from $\mathcal{C}$ to a future unlabeled observation x.
2. Assess the uncertainty in each classification (i.e., the probability of misclassification).
3. Understand the roles of the different predictors among $X = (X_1, X_2, \ldots, X_p)$.

Classification problem

In place of MSE, we now use:

Error rate

$$\text{Error rate} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{f}(x_i))$$

where $\hat{f}(x_i)$ is the predicted class label
and $I(y_i \neq \hat{f}(x_i))$ is an indicator function.

- That is, the error rate is the fraction of misclassifications.
- The training error rate is misleading (too small).
- We want to minimize the test error rate: $\mathrm{E}(I(y_0 \neq \hat{y}_0))$

Classification problem

In place of MSE, we now use:

Error rate

$$\text{Error rate} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{f}(x_i))$$

where $\hat{f}(x_i)$ is the predicted class label and $I(y_i \neq \hat{f}(x_i))$ is an indicator function.

- That is, the error rate is the fraction of misclassifications.
- The training error rate is misleading (too small).
- We want to minimize the test error rate: $\mathrm{E}(I(y_0 \neq \hat{y}_0))$

Optimal classifier

Suppose the K elements in $\mathcal{C}$ are numbered $1, 2, \ldots, K$. Let

$$p_k(x) = \Pr(Y = k \mid X = x), \qquad k = 1, 2, \ldots, K.$$

These are the conditional class probabilities at x.

Then the Bayes classifier at x is

$$C(x) = j \quad \text{if } p_j(x) = \max\{p_1(x), p_2(x), \ldots, p_K(x)\}$$

- This gives the minimum average test error rate.
- It is an "oracle predictor" because we do not usually know $p_K(x)$.

Optimal classifier

Suppose the K elements in $\mathcal{C}$ are numbered $1, 2, \ldots, K$. Let

$$p_k(x) = \Pr(Y = k \mid X = x), \qquad k = 1, 2, \ldots, K.$$

These are the conditional class probabilities at x.
Then the Bayes classifier at x is

$$C(x) = j \quad \text{if } p_j(x) = \max\{p_1(x), p_2(x), \ldots, p_K(x)\}$$

- This gives the minimum average test error rate.
- It is an "oracle predictor" because we do not usually know $p_K(x)$.

Optimal classifier

Suppose the K elements in $\mathcal{C}$ are numbered $1, 2, \ldots, K$. Let

$$p_k(x) = \Pr(Y = k \mid X = x), \qquad k = 1, 2, \ldots, K.$$

These are the conditional class probabilities at x.
Then the Bayes classifier at x is

$$C(x) = j \quad \text{if } p_j(x) = \max\{p_1(x), p_2(x), \ldots, p_K(x)\}$$

- This gives the minimum average test error rate.
- It is an “oracle predictor” because we do not usually know $p_k(x)$.

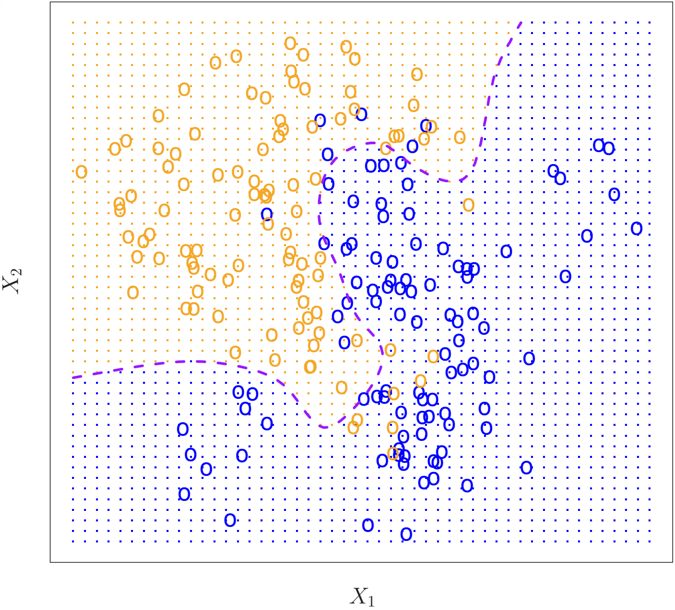
Bayes error rate

Bayes error rate

$$1 - \mathrm{E}\left(\max_j \Pr(Y = j|X)\right)$$

- The “Bayes error rate” is the lowest possible error rate that could be achieved if we knew exactly the “true” probability distribution of the data.
- It is analagous to the “irreducible error” in regression.
- On test data, no classifier can get lower error rates than the Bayes error rate.
- In reality, the Bayes error rate is not known exactly.

Bayes optimal classifier



k-Nearest Neighbours

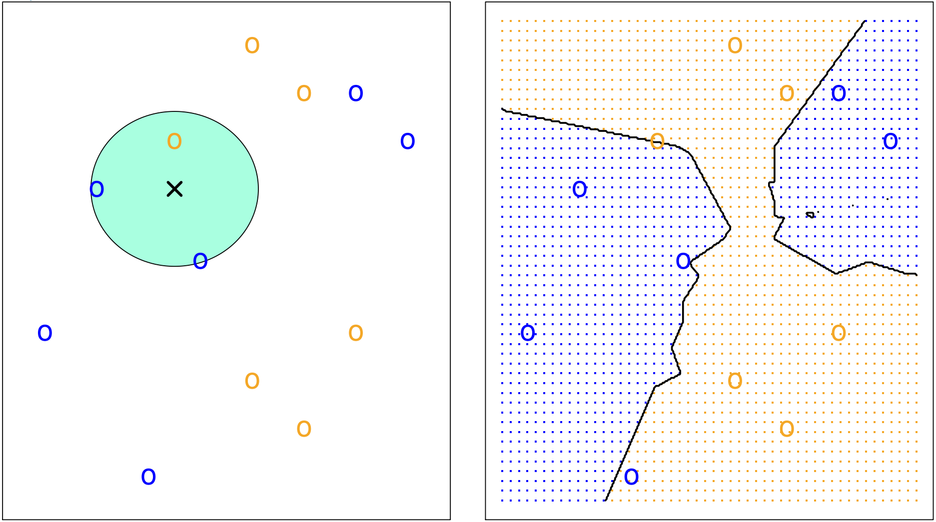
One of the simplest classifiers. Given a test observation x_0:

- Find the K nearest points to x_0 in the training data: $\mathcal{N}_0$.
- Estimate conditional probabilities

$$\Pr(Y = j \mid X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j).$$

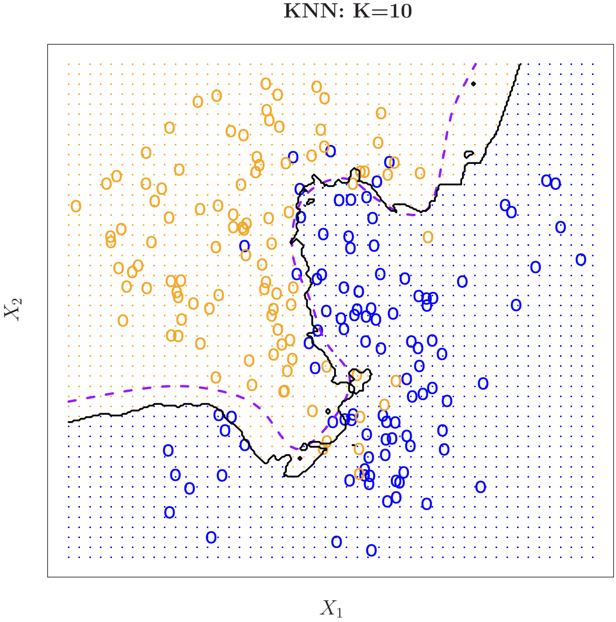
- Apply Bayes rule and classify x_0 to class with largest probability.

kNN Classifier



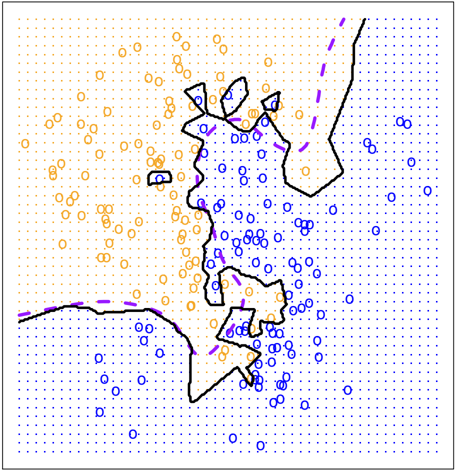
$K = 3$.

kNN Classifier

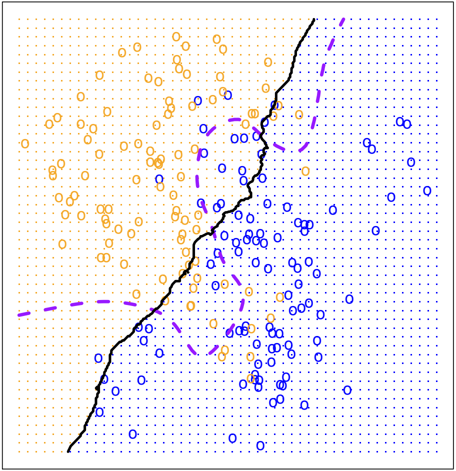


kNN Classifier

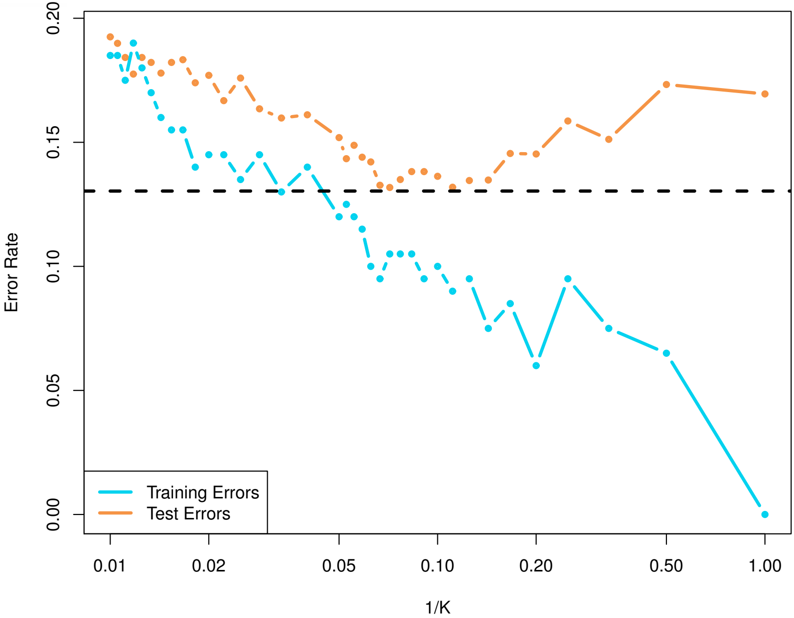
KNN: K=1



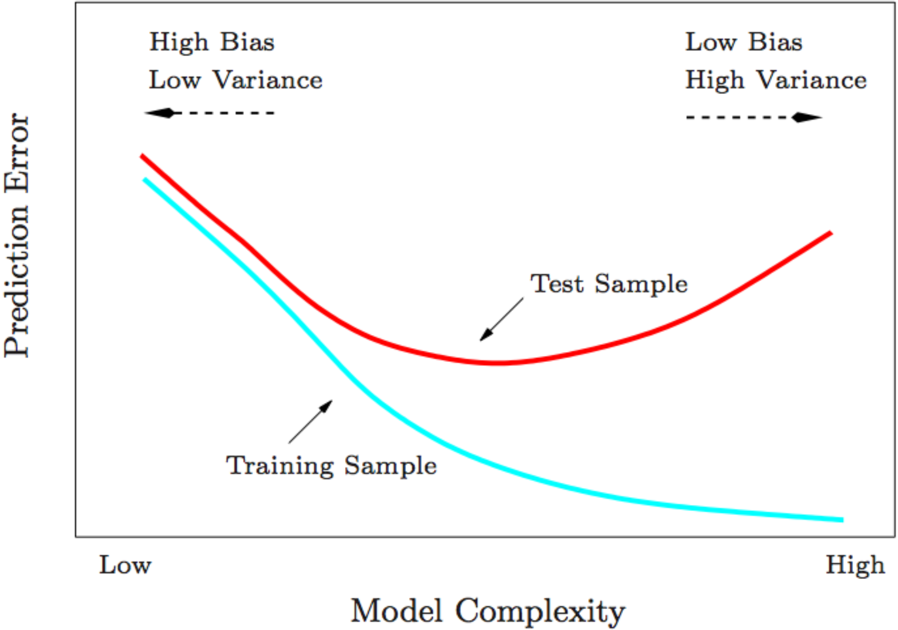
KNN: K=100



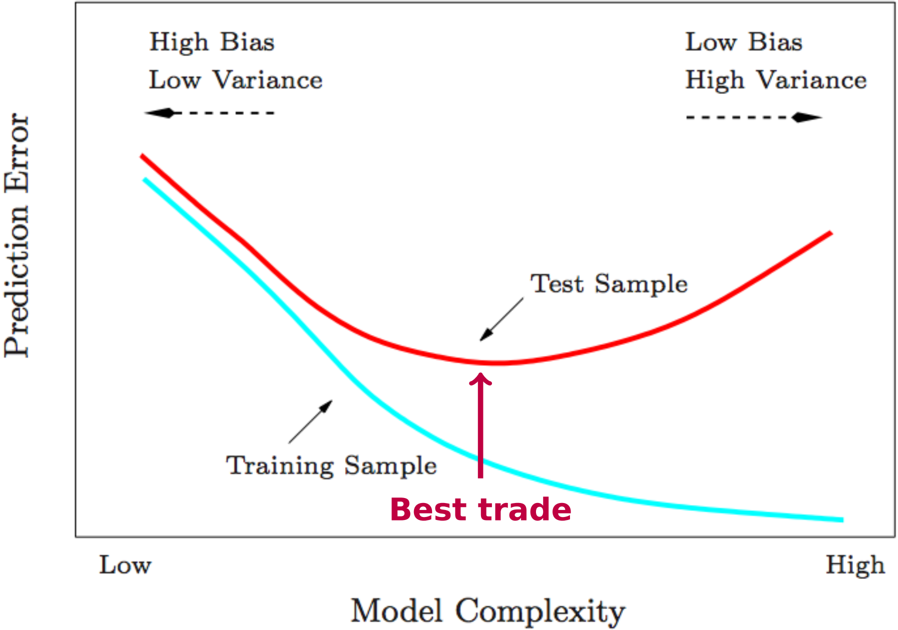
kNN Classifier



A fundamental picture



A fundamental picture



A fundamental picture