Simple model $\rightarrow$ high bias
  e.g. just linear

Complex model: 'fit anything' 
  low bias 
  high variance

Squared error

\[ l = (y - \hat{y})^2 \]

\[ y \leftarrow \text{true target} \]
\[ \hat{y} \leftarrow \text{predicted value} \]

Expected loss

\[ \mathbb{E}_{x}[(y - \hat{y})^2] \]
\[ \mathbb{E}[(y - \mu_y)^2] \]

(\(x, y\) given)
\[ \hat{y} = M(x) \]
Bias-Variance tradeoff

\[ \mu_\delta = \mathbb{E}[\delta] \]

\[ \mu_\epsilon = \mathbb{E}[\epsilon] \]

\[ \mu_\tau = \mathbb{E}[\tau] \]

\[ \mathbb{E}[(\hat{y} - \mu_y)^2] \] 
average bias term

\[ \mathbb{E}[(\delta - \mu_\delta)^2] \] 
average variance term

Ensemble methods

Bootstrap Aggregation (Bagging)

\[ \Rightarrow \text{ average from multiple models} \to \text{ reducing variance} \]

Bagging \((D, \text{Alg})\)

\[ \text{for } i = 1 \ldots k \]

\[ D_i = \text{bootstrap sample } (D) \]

\[ \mu_i = \text{Alg} (D_i) \]

\[ \mu_1, \mu_2, \ldots, \mu_k \]

\[ k \text{ different models} \]

\[ \text{given any point } \boldsymbol{x} \]
\[ \hat{y}_1 = M_1(\hat{x}) \]
\[ \hat{y}_2 = M_2(\hat{x}) \]
\[ \vdots \]
\[ \hat{y}_k = M_k(\hat{x}) \]

**Classification**: majority vote \( \left\{ \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_k \right\} \)

\[ H \quad H \quad L \quad L \quad L \]

**Regression**: \[ E[\hat{y}] = \frac{1}{k} (\hat{y}_1 + \hat{y}_2 + \ldots + \hat{y}_k) \]

Reduce variance: better applied to a more complex model (low bias)

will help lower variance

**Random Forests**: Ensemble of decision trees on bootstrapped samples

Sample the attributes to split on at every node

\( (A_1, A_2, \ldots, A_d) \)

Sample \( m \ll d \) attributes

\( A_1, A_2, \ldots \)

**Boosting**: reduce bias via ensemble
Adaptive Boosting (AdaBoost)

$\overrightarrow{w} = \left( \frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n} \right)$

prob vector $w_i = \text{prob that point } x_i \text{ will be sampled}$

for $i = 1, \ldots, k$

$D_i = \text{sample } D \text{ according to } \overrightarrow{w} \text{ of size } n$

$M_i = \text{Alg}(D_i)$

$\varepsilon_i = \text{error on } D$

$\alpha_i = \ln \left( \frac{1}{\varepsilon_i} - 1 \right) \equiv \text{score of classifier } M_i$

$\forall j = 1, \ldots, n$

if $M_i$ makes no mistake on $x_j$

then $w_j$ stays the same

else: if $M_i$ makes a mistake

$w_j = w_j \cdot \exp(\alpha_i)$

$= w_j \cdot \exp \left( \ln \left( \frac{1}{\varepsilon_i} - 1 \right) \right)$

$= w_j \cdot \left( \frac{1}{\varepsilon_i} - 1 \right)$

$\varepsilon_i \leq 0.5$

$\ln \left( \frac{1}{\varepsilon_i} - 1 \right) = \ln(2-1) = \ln(1) = 0$

$\varepsilon_i \leq 0.25 = 1/4$

$\ln \left( \frac{1}{1/4} - 1 \right) = \ln(4-1) = \ln(3)$

$\varepsilon_i \leq \frac{1}{100}$

$\alpha_i = \ln(98)$
\[ L = w_j \cdot \left( \frac{1}{2} - 1 \right) \]

\[ w_j = \frac{3}{N \cdot y_j} \]

D_2: these points will have higher prob

\[ h_1(\hat{x}) \Rightarrow o \cdot \alpha_1 \]
\[ h_2(\hat{x}) \Rightarrow o \cdot \alpha_2 \]

\[ \sqrt{\text{final prediction: not simple majority voting}} \]

\[ \text{Weighted vote} \]

Unsupervised Methods

Pattern Mining

clustering.

Pattern Mining
Frequently occurring trends

\( \Rightarrow \) co-occurrence patterns

\( \Rightarrow \) Itemset mining

**Transaction database**

\[ I = \{ \text{set of items} \} = \{ \{ A, B, C, D, E \} \} \]

\[ T = \{ \text{subset of items} \} = \begin{array}{c}
1. \{ A, B, D, E \} \\
2. \{ B, C, E \} \\
3. \{ A, B, D, E \} \\
4. \{ A, B, C \} \\
5. \{ A, B, C, D, E \} \\
6. \{ B, C, D, E \} \\
\end{array} \]

\[ X = \{ A, B \} = AB \]

\[ \text{Sup}(X) = Y \]

\[ T_i \subseteq I \]

Q: What are the frequently occurring subsets?

\[ \text{Sup}(X) = \text{absolute suppor}(X) = \text{count}(X) \]

\[ P(X) = \frac{\text{Sup}(X)}{n} \]

prob of \( X \)

(joint prod of all the item in \( X \))

relative suppor \( \equiv \) \( P(X) \)

\[ \text{minsup} \equiv \text{minimum support threshold} \]

\( \Rightarrow \) absolute

3

\( \Rightarrow \) relative

0.5
Q: given minsup threshold, find all frequent patterns
le. all X \subseteq I
such that \text{sup}(X) \geq \text{minsup}

F = \text{set of all frequent itemsets}