Decision Trees (for classification)

- Full binary tree
  Every vertex has either 0, 1, or 2 children
  or is a leaf
  - Label of internal nodes:
    Feature compared with threshold
  - Label of leaves: +1, -1.

Decision Tree $(S, k)$

If $|S| \leq k$:

return Leaf (label = argmax $\hat{P}_S(a)$)

else:

for each $j, \theta$:

\[
\begin{align*}
S_{N,j} &= \{ (x,y) \mid x_j < \theta \} \\
S_{Y,j} &= \{ (x,y) \mid x_j \geq \theta \} \\
C(j, \theta) &= \frac{\sum_{i \in S_{N,j}} (y_i - y^{(i)})^2}{\sum_{i \in S_{Y,j}} (y_i - y^{(i)})^2} + \frac{S_{N,j} - 1}{S_{Y,j}}
\end{align*}
\]

$j^*, \theta^* = \text{argmin} C(j, \theta)$

return Node (label = "$j^* > \theta^*$", Decision Tree $(S_{N,j}, k)$, Decision Tree $(S_{Y,j}, k)$)

\[
\begin{align*}
\hat{P}_S(a) &= \frac{1}{|S|} \sum_{i \in S} y^{(i)} \\
\hat{P}_S^{*}(a) &= \frac{1}{|S|} \sum_{i \in S} \hat{P}_S(a)
\end{align*}
\]
Decision Trees are prone to overfitting

**Bagging**: Boosting Aggregation

- Given training set $S$, construct training sets $S^{(1)}, S^{(2)} \ldots S^{(k)}$ by picking (with replacement) $m$ examples from $S$.
- Build decision tree $T^{(1)}, T^{(2)} \ldots T^{(k)}$ on $S^{(1)}, S^{(2)} \ldots S^{(k)}$.

**Aggregation**
- On a new example $(x)$
  \[
  \hat{y} = \text{maj}_k T^{(i)}(x) \quad \text{(classification)}
  \]
  \[
  \hat{y} = \frac{1}{k} \sum_{i=1}^{k} T^{(i)}(x) \quad \text{(regression)}
  \]

**Random Forests**
- Given training set $S$, construct training sets $S^{(1)}, \ldots S^{(k)}$ by picking (with replacement) $m$ examples from $S$.
- Build a decision tree $T^{(i)}$ on $S^{(i)}$ as follows.
  - At each stage of the decision tree construction, pick a random subset of features $I$, and you use one of the features in $I$ to split.
- On a new example $x$.
  "Aggregate" the outputs of $T^{(i)}(x)$. 
k-Nearest Neighbors:

No hypothesis constructed from the training set.

On a new example $x$, $S = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$.
- Compute permutation of $S$ say $(x^{\pi(1)}, y^{\pi(1)}), \ldots, (x^{\pi(n)}, y^{\pi(n)})$ such that
  \[d(x, x^{\pi(i)}) \leq d(x, x^{\pi(i+1)}) - \sqrt{\sum_{j=1}^{k} (y^{\pi(i)} - y^{\pi(i+1)})^2}].
- Output $y = \text{maj}(y^{\pi(1)}, y^{\pi(2)}, \ldots, y^{\pi(k)})$ (classification)
  \[y = \frac{1}{k} \sum_{i=1}^{k} y^{\pi(i)} \quad (\text{regression})\]

Examples that are close by have outputs that are close.

$c$-Lipschitz: $c d(x^{(i)}, x^{(j)}) \geq d(y^{(i)}, y^{(j)})$