Decision trees

Decision trees

Directly optimize tree structure for good classification.

A **decision tree** is a function $f \colon \mathcal{X} \to \mathcal{Y}$, represented by a binary tree in which:

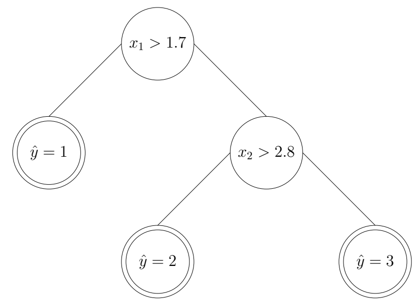
- Each **tree node** is associated with a splitting rule $g \colon \mathcal{X} \to \{0, 1\}$.
- Each **leaf node** is associated with a label $y \in \mathcal{Y}$.

When $\mathcal{X} = \mathbb{R}^d$, typically only consider splitting rules of the form

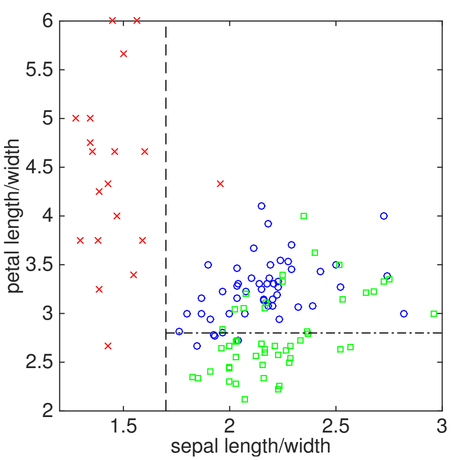
$$g(\boldsymbol{x}) \;=\; \mathbb{1}\{x_i > t\}$$

for some $i \in [d]$ and $t \in \mathbb{R}$.
Called *axis-aligned* or *coordinate* splits.

(Notation: $[d] := \{1, 2, \ldots, d\}$)

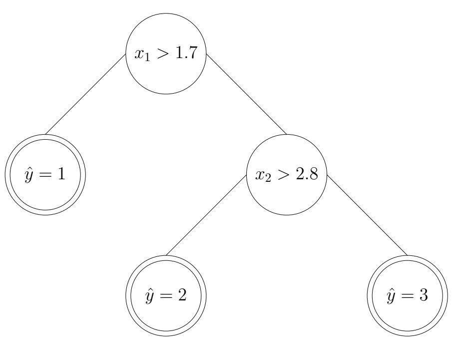


Decision tree example

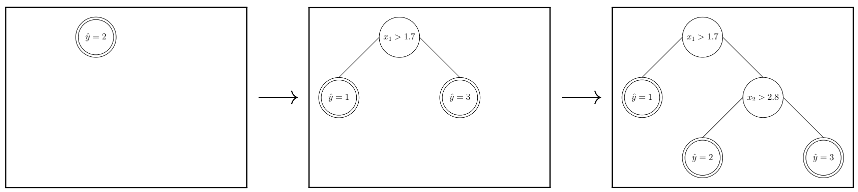


Classifying irises by sepal and petal measurements

- $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{1, 2, 3\}$
- x_1 = ratio of sepal length to width
- x_2 = ratio of petal length to width



Basic decision tree learning algorithm



Basic "top-down" greedy algorithm

- Initially, tree is a single leaf node containing all (training) data.
- Loop:
  - Pick the leaf ℓ and rule h that maximally reduces uncertainty.
  - Split data in ℓ using h, and grow tree accordingly.

  . . . until some stopping criterion is satisfied.

[**Label of a leaf** is the **plurality label** among the data contained in the leaf.]

Notions of uncertainty

Notions of uncertainty: binary case ($\mathcal{Y} = \{0, 1\}$)

Suppose in a set of examples $S \subseteq \mathcal{X} \times \{0, 1\}$, a p fraction are labeled as 1.

1. **Classification error**:

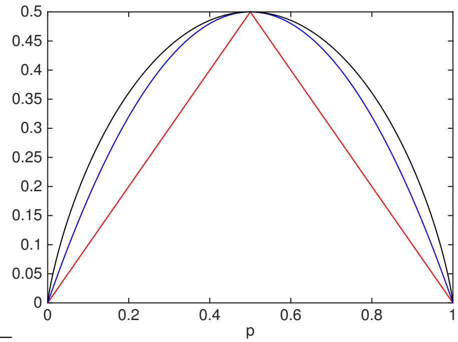
$$u(S) := \min\{p,\ 1 - p\}$$

2. **Gini index**:

$$u(S) := 2p(1 - p)$$

3. **Entropy**:

$$u(S) := p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$$



Gini index and entropy (after some rescaling) are concave upper-bounds on classification error.

Notions of uncertainty

Notions of uncertainty: general case ($\mathcal{Y} = \{1, 2, \dots, K\}$)

Suppose in $S \subseteq \mathcal{X} \times \mathcal{Y}$, a p_y fraction are labeled as y (for each $y \in \mathcal{Y}$).

1. **Classification error**:

$$u(S) := 1 - \max_{y \in \mathcal{Y}} p_y$$

2. **Gini index**:

$$u(S) := 1 - \sum_{y \in \mathcal{Y}} p_y^2$$

3. **Entropy**:

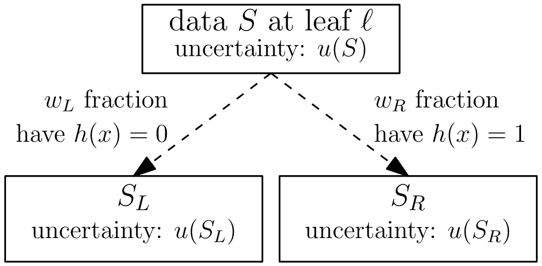
$$u(S) := \sum_{y \in \mathcal{Y}} p_y \log \frac{1}{p_y}$$

Each is *maximized* when $p_y = 1/K$ for all $y \in \mathcal{Y}$ (i.e., equal numbers of each label in S).

Each is *minimized* when $p_y = 1$ for a single label $y \in \mathcal{Y}$ (so S is **pure** in label).

Uncertainty reduction

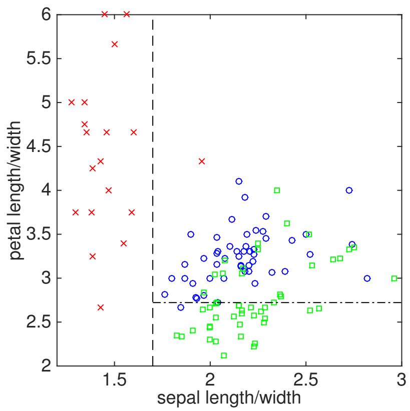
Suppose the data S at a leaf ℓ is split by a rule h into S_L and S_R, where $w_L := |S_L|/|S|$ and $w_R := |S_R|/|S|$.

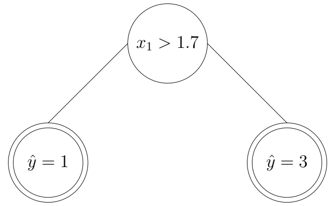


The **reduction in uncertainty** from using rule h at leaf ℓ is

$$u(S) - \Big(w_L \cdot u(S_L) + w_R \cdot u(S_R)\Big).$$

Uncertainty reduction





One leaf (with $\hat{y} = 1$) already has zero uncertainty (a pure leaf).

Other leaf (with $\hat{y} = 3$) has Gini index

$$u(S) = 1 - \left(\frac{1}{101}\right)^2 - \left(\frac{50}{101}\right)^2 - \left(\frac{50}{101}\right)^2 = 0.5098\,.$$

Split S with $\mathbb{1}\{x_2 > 2.7222\}$ to S_L, S_R:

$$u(S_L) = 1 - \left(\tfrac{0}{30}\right)^2 - \left(\tfrac{1}{30}\right)^2 - \left(\tfrac{29}{30}\right)^2 = 0.0605,$$

$$u(S_R) = 1 - \left(\tfrac{1}{71}\right)^2 - \left(\tfrac{49}{71}\right)^2 - \left(\tfrac{21}{71}\right)^2 = 0.4197.$$

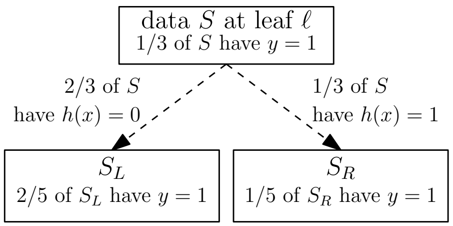
Reduction in uncertainty:

$$0.5098 - \left(\frac{30}{101} \cdot 0.0605 + \frac{71}{101} \cdot 0.4197\right) = 0.2039\,.$$

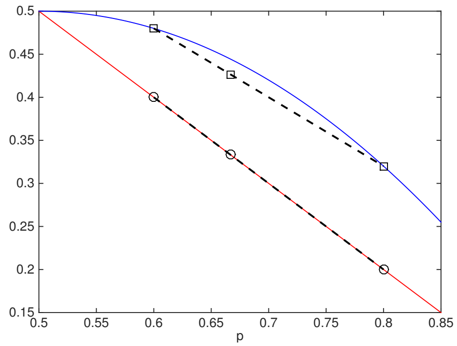
Comparing notions of uncertainty

It is possible to have a splitting rule h with

- zero reduction in classification error, but
- non-zero reduction in Gini index or entropy.



$(\mathcal{Y} = \{0, 1\})$



Reduction in classification error: $\frac{1}{3} - \left(\frac{2}{3} \cdot \frac{2}{5} + \frac{1}{3} \cdot \frac{1}{5}\right) = 0$

Reduction in Gini index: $\frac{4}{9} - \left(\frac{2}{3} \cdot \frac{12}{25} + \frac{1}{3} \cdot \frac{8}{25}\right) = \frac{4}{225}$

Aside: reduction in entropy

The (Shannon) **entropy** of $\mathcal{Z}$-valued random variable Z with is

$$H(Z) \ := \ \sum_{z \in \mathcal{Z}} \mathbb{P}(Z = z) \log \frac{1}{\mathbb{P}(Z = z)}.$$

The **conditional entropy** of Z given a $\mathcal{W}$-valued random variable W is

$$H(Z|W) \ := \ \sum_{w \in \mathcal{W}} \mathbb{P}(W = w) \cdot H(Z|W = w).$$

(Weighted average of entropies of random variables of form $Z|W = w$.)

Think of (X, Y) as random pair taking value $(x, y) \in S$ with probability $1/|S|$.
Using entropy as uncertainty measure, $u(S) = H(Y)$.

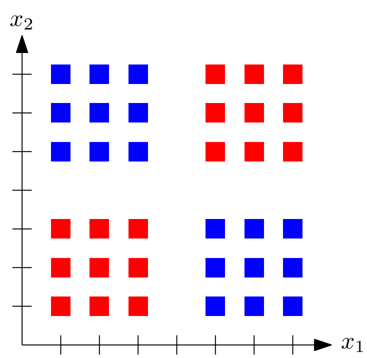
Reduction in uncertainty after a split $g \colon \mathcal{X} \to \{0, 1\}$ is

$$H(Y) - \Big(\mathbb{P}(g(X) = 0) \cdot H(Y|g(X) = 0) + \mathbb{P}(g(X) = 1) \cdot H(Y|g(X) = 1)\Big)$$
$$= H(Y) - H(Y|g(X)).$$

This is called **mutual information** between Y and $g(X)$.

Limitations of uncertainty notions

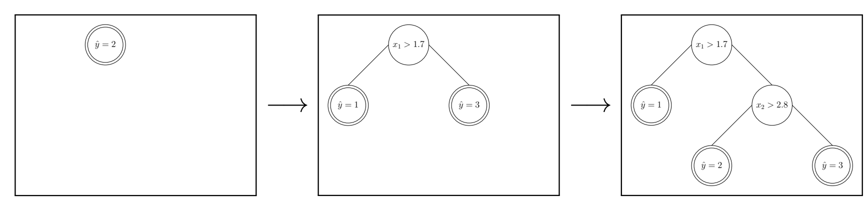
Suppose $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{Y} = \{\text{red}, \text{blue}\}$, and the data is as follows:



Every split of the form $\mathbb{1}\{x_i > t\}$ provides no reduction in uncertainty (whether based on classification error, Gini index, or entropy).

Upshot:
Zero reduction in uncertainty may not be a good stopping condition.

Basic decision tree learning algorithm



Basic "top-down" greedy algorithm

- Initially, tree is a single leaf node containing all (training) data.
- Loop:
  - Pick the leaf ℓ and rule h that maximally reduces uncertainty.
  - Split data in ℓ using h, and grow tree accordingly.

. . . until some stopping criterion is satisfied.

[**Label of a leaf** is the **plurality label** among the data contained in the leaf.]

Stopping criterion

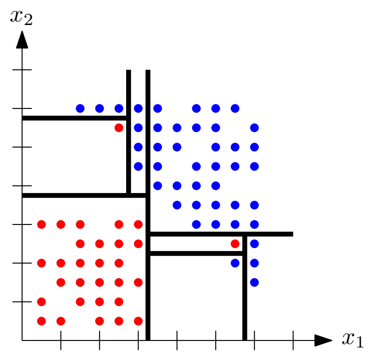
Many alternatives; two common choices are:

1. Stop when the **tree reaches a pre-specified size**.

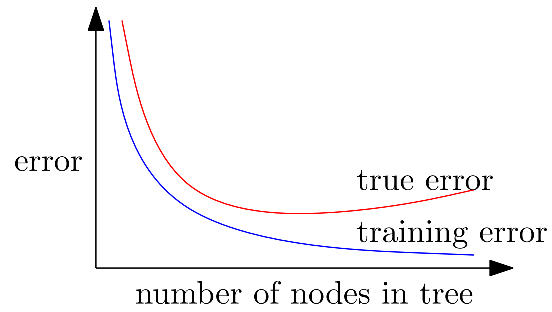
   Involves setting additional "tuning parameters" (similar to k in k-NN).

2. Stop when **every leaf is pure**. (More common.)

   Serious danger of **overfitting** spurious structure due to sampling.



Overfitting



- Training error **goes to zero** as the number of nodes in the tree increases.
- True error decreases initially, but eventually **increases due to overfitting**.

What can be done about overfitting?

Preventing overfitting

Split training data S into two parts, S' and S'':

- Use first part S' to **grow the tree until all leaves are pure**.
- Use second part S'' to **choose a good pruning of the tree**.

Pruning algorithm

Loop:

- Replace any tree node by a leaf node if it improves the error on S''.

. . . until no more such improvements possible.

This can be done efficiently using **dynamic programming** (bottom-up traversal of the tree).

Independence of S' and S'' make it unlikely for spurious structures in each to perfectly align.

Example: Spam filtering

Data

- 4601 e-mail messages, 39.4% are spam.
- $\mathcal{Y} = \{\text{spam}, \text{not spam}\}$
- E-mails represented by 57 features:
  - 48: percentange of e-mail words that is specific word (e.g., "free", "business")
  - 6: percentage of e-mail characters that is specific character (e.g., "!").
  - 3: other features (e.g., average length of ALL-CAPS words).

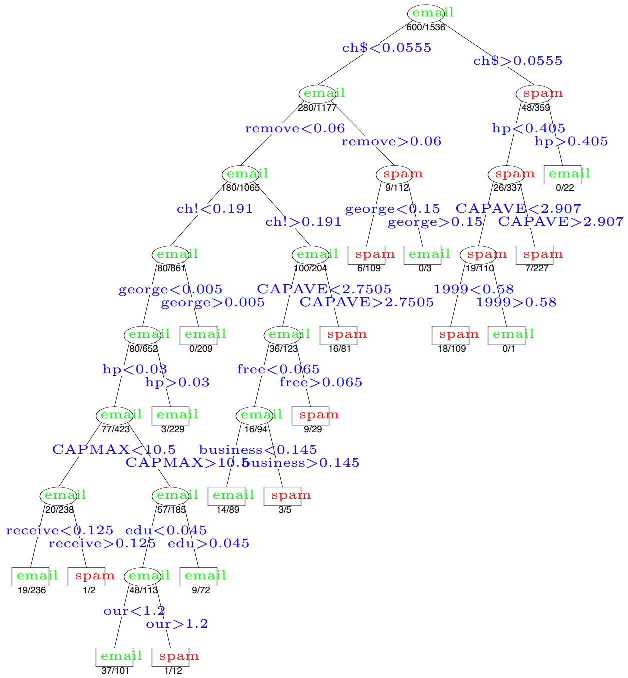
Results

Using variant of greedy algorithm to grow tree; prune tree using validation set.

Chosen tree has just **17 leaves**. Test error is 9.3%.

	$\hat{y}$ = not spam	$\hat{y}$ = spam
y = not spam	57.3%	4.0%
y = spam	5.3%	33.4%

Example: Spam filtering



Final remarks

- Decision trees are very flexible classifiers (like NN).
  - Certain greedy strategies for training decision trees are **consistent**.
  - But also **very prone to overfitting** in most basic form.
  - (NP-hard to find smallest decision tree consistent with data.)
- Current theoretical understanding of (greedy) decision tree learning:
  - As fitting a **non-parametric model** (like NN).
  - As **meta-algorithm** for combining classifiers (splitting rules).

Key takeaways

1. Structure of decision tree classifiers.
2. Greedy learning algorithm based on notions of uncertainty; limitations of the greedy algorithm.
3. High-level idea of overfitting and ways to deal with it.