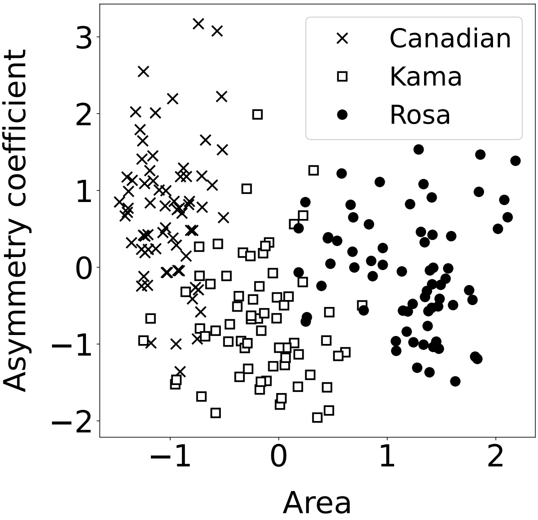
Classification Trees

Probability and Statistics for Data Science

Carlos Fernandez-Granda

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Classification



Classification

Data: (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)

Each feature x_i is a d-dimensional vector

The label y_i indicates the class (e.g. *Canadian*, *Kama*, or *Rosa*)

Goal: Assign class to new data

Probabilistic modeling

Model features as random vector $\tilde{x}$ and label as random variable $\tilde{y}$

For new data vector x:

$$\hat{y} := \arg \max_{y \in \{1,2,\ldots,c\}} p_{\tilde{y} \,|\, \tilde{x}}(y \,|\, x)$$

Is classification easy? No, due to curse of dimensionality!

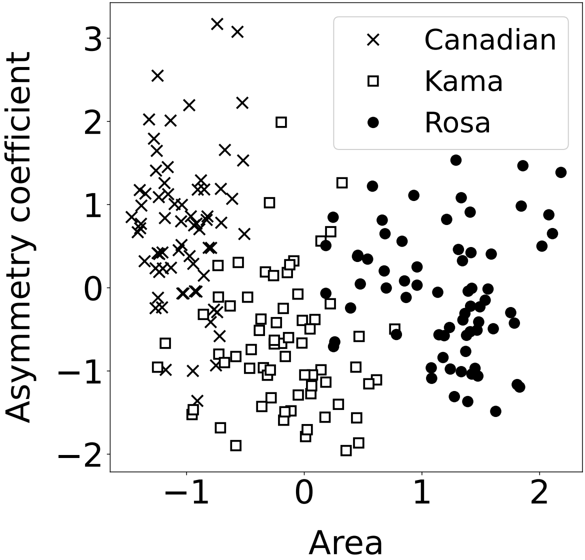
Discriminative classification

Goal: Approximate $p_{\tilde{y} \mid \tilde{x}}(k \mid x)$ for $1 \leq k \leq c$

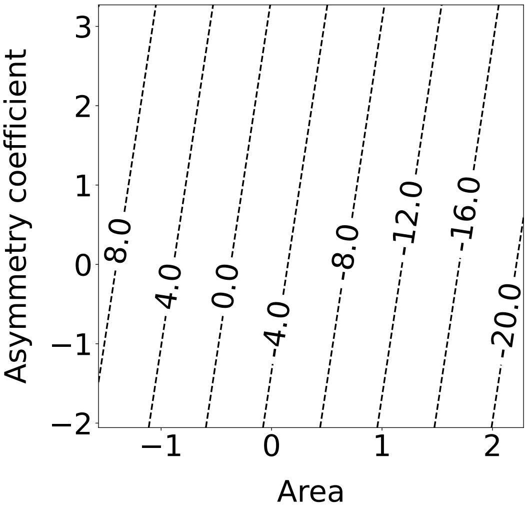
Logistic and softmax regression:

Linear function of features mapped to probabilities

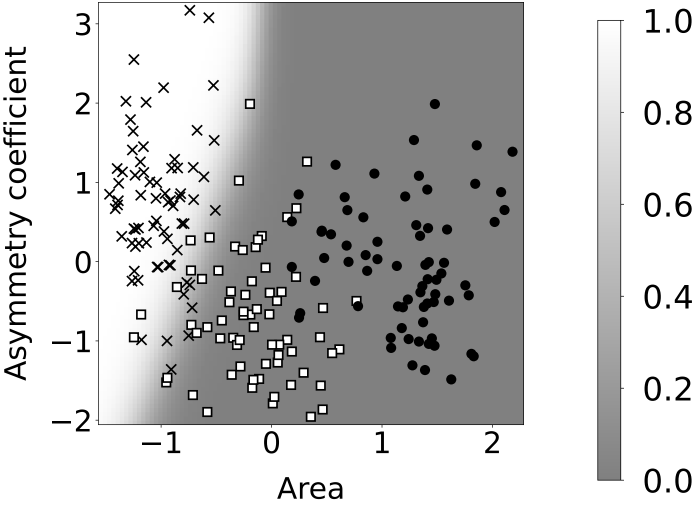
Wheat varieties



Canadian: $-7.7\,a + 0.9\,c - 2.9$



Canadian: Estimated probability



Goal

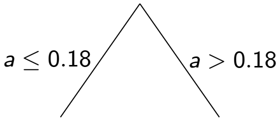
Model nonlinear relationship between features and response

Idea: Build a classification tree, in the spirit of regression trees

1. Build binary tree that partitions the feature space into disjoint regions
2. Assign constant probabilities $p_{\tilde{y} \mid \tilde{x}}(k \mid x)$ for $1 \leq k \leq c$ in each region

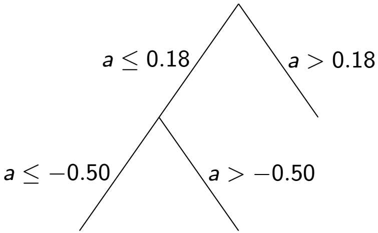
Classification tree for wheat varieties

Features: area (a), asymmetric coefficient (c)



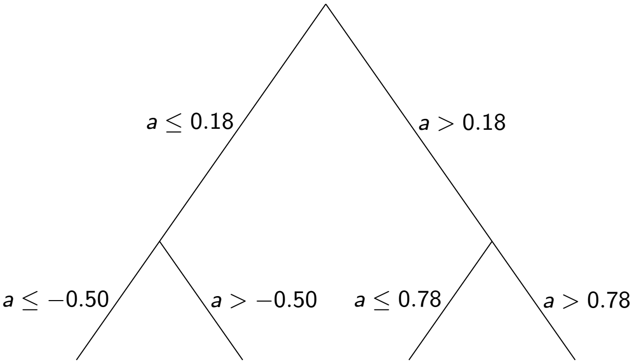
Classification tree for wheat varieties

Features: area (a), asymmetric coefficient (c)



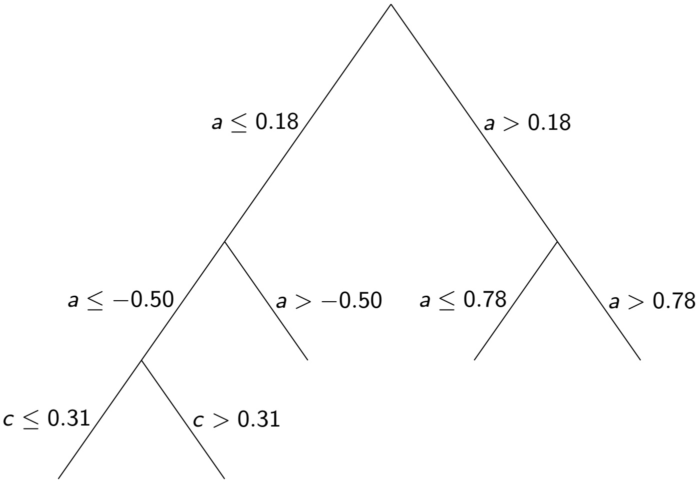
Classification tree for wheat varieties

Features: area (a), asymmetric coefficient (c)



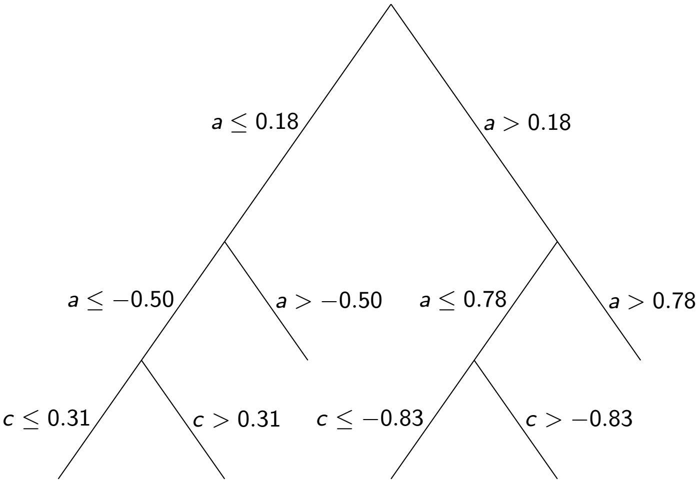
Classification tree for wheat varieties

Features: area (a), asymmetric coefficient (c)

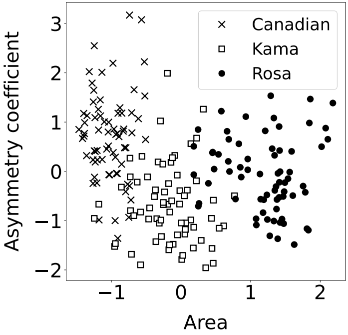


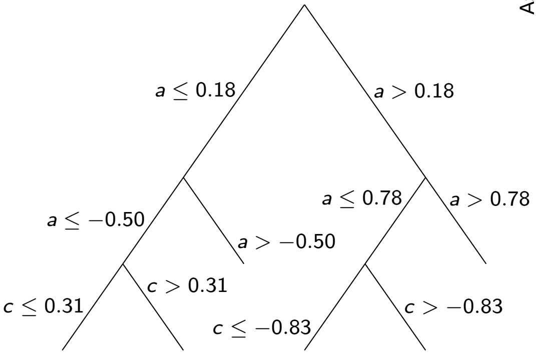
Classification tree for wheat varieties

Features: area (a), asymmetric coefficient (c)



Classification tree for wheat varieties





Two key questions

How to compute constant probability estimates?

How to build the tree?

Constant probability estimate?

Consider the n_R feature-response pairs (x_i, y_i) in region R

Goal: Choose constant probability estimate θ

We maximize conditional likelihood of labels given features

Likelihood

We model ith feature and label as random variables $\tilde{x}_i$ and $\tilde{y}_i$

Assumption 1:
Labels are conditionally independent given the features

Assumption 2:
$\tilde{y}_i$ is conditionally independent from $\{\tilde{x}_m\}_{m \neq i}$ given $\tilde{x}_i$

$$\begin{aligned}
\mathcal{L}_{XY}(\theta) &:= \mathrm{P}\left(\tilde{y}_1 = y_1, ..., \tilde{y}_n = y_n \,|\, \tilde{x}_1 = x_1, ..., \tilde{x}_n = x_n\right) \\
&= \prod_{i=1}^{n} \mathrm{P}\left(\tilde{y}_i = y_i \,|\, \tilde{x}_1 = x_1, ..., \tilde{x}_n = x_n\right) \\
&= \prod_{i=1}^{n} \mathrm{P}\left(\tilde{y}_i = y_i \,|\, \tilde{x}_i = x_i\right) \\
&= \prod_{k=1}^{c} \prod_{\{i: y_i = k\}} \theta[k]
\end{aligned}$$

Maximum likelihood

$$\mathcal{L}_{XY}(\theta) = \prod_{k=1}^{c} \prod_{\{i:y_i=k\}} \theta[k]$$

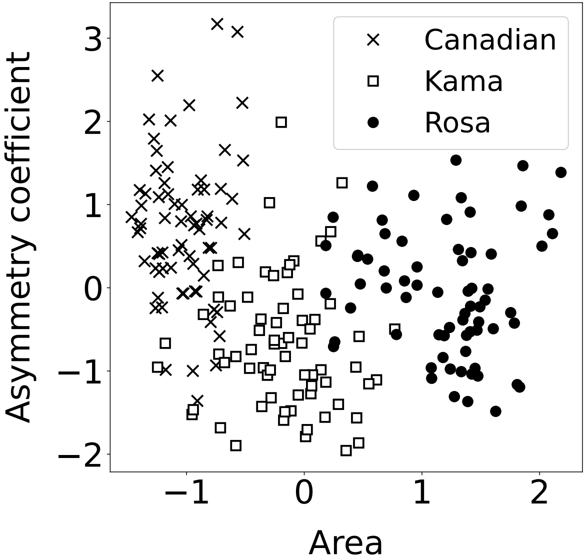
For $c := 2$? Bernoulli likelihood

$$\theta_{\text{ML}}[1] = \frac{n_1}{n}$$
$$\theta_{\text{ML}}[2] = \frac{n_2}{n}$$

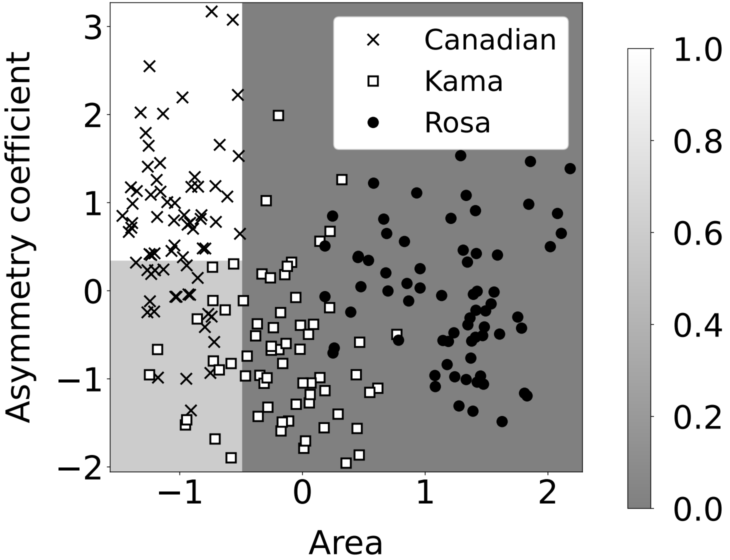
In general, for an arbitrary number of classes c

$$\theta_{\text{ML}}[k] = \frac{n_k}{n} \qquad 1 \leq k \leq c$$

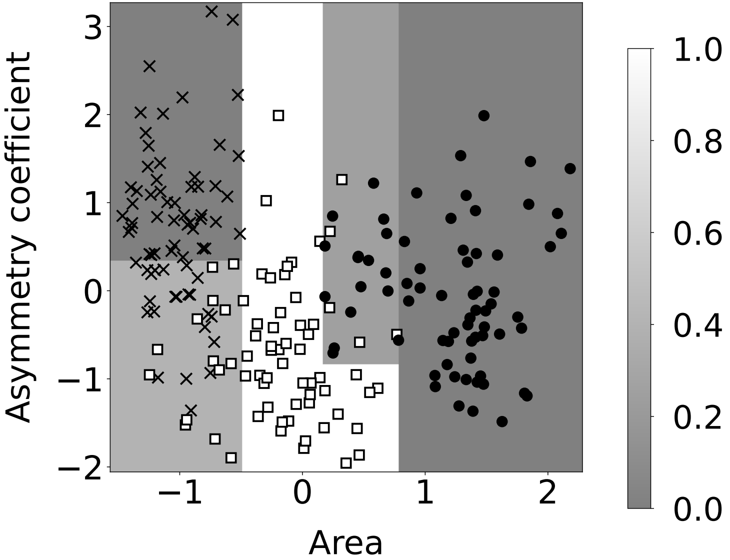
Wheat varieties



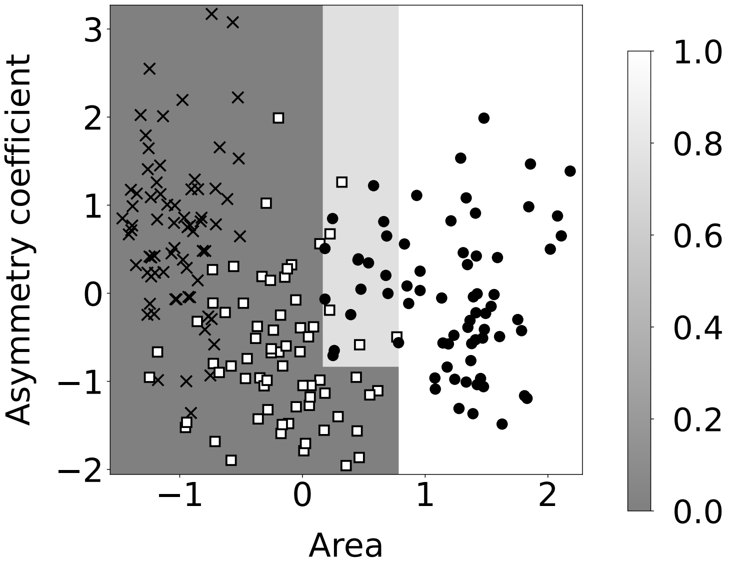
Canadian: Estimated probability



Kama: Estimated probability



Rosa: Estimated probability



How to build the tree?

Idea: Choose tree with maximum likelihood

Problem: Intractable, too many possible trees...

Solution: Recursive binary splitting, greedily choose bifurcations

Log-likelihood

We model ith feature and label as random variables $\tilde{x}_i$ and $\tilde{y}_i$

Assumption 1:
Labels are conditionally independent given the features

Assumption 2:
$\tilde{y}_i$ is conditionally independent from $\{\tilde{x}_m\}_{m \neq i}$ given $\tilde{x}_i$

$$\mathcal{L}_{XY}(\theta) = \prod_{i=1}^{n} \mathrm{P}\left(\tilde{y}_i = y_i \,|\, \tilde{x}_i = x_i\right)$$

For tree with regions $\mathcal{R} := \{R_1, \ldots, R_m\}$

$$\mathcal{L}_{XY}(\mathcal{R}) = \prod_{k=1}^{c} \prod_{\{i : y_i = k\}} \theta_{R(x_i)}[k]$$

$$\log \mathcal{L}_{XY}(\mathcal{R}) = \sum_{k=1}^{c} \sum_{\{i : y_i = k\}} \log \theta_{R(x_i)}[k]$$

Likelihood-based splitting

Choose split that most increases log-likelihood

$$\log \mathcal{L}_{XY}(\mathcal{R}) = \sum_{k=1}^{c} \sum_{\{i : y_i = k\}} \log \theta_{R(x_i)}[k]$$

If region R is split into A and B

$$\triangle \mathcal{L}_{XY} = \sum_{k=1}^{c} \left(n_A^{[k]} \log \theta_A[k] + n_B^{[k]} \log \theta_B[k] - n_R^{[k]} \log \theta_R[k] \right)$$

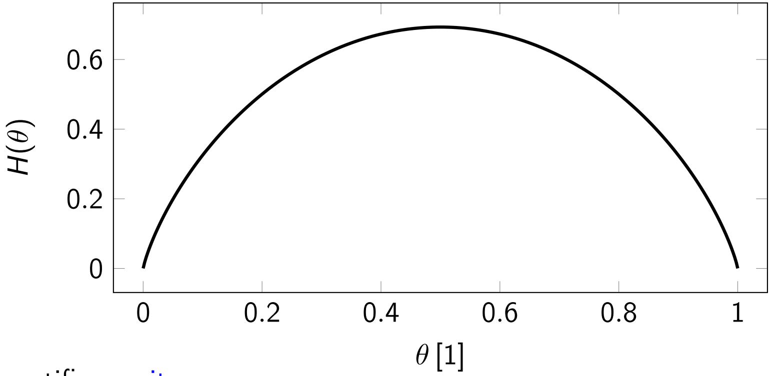
Entropy

The entropy of a pmf θ with c entries is

$$H(\theta) := -\sum_{k=1}^{c} \theta[k] \log \theta[k]$$

Metric to quantify information content

Entropy for $c := 2$



Quantifies purity

Low entropy -> One class dominates -> Good for classification!

Alternative: Gini index

$$G(\theta) := -\sum_{k=1}^{c} \theta[k](1 - \theta[k])$$

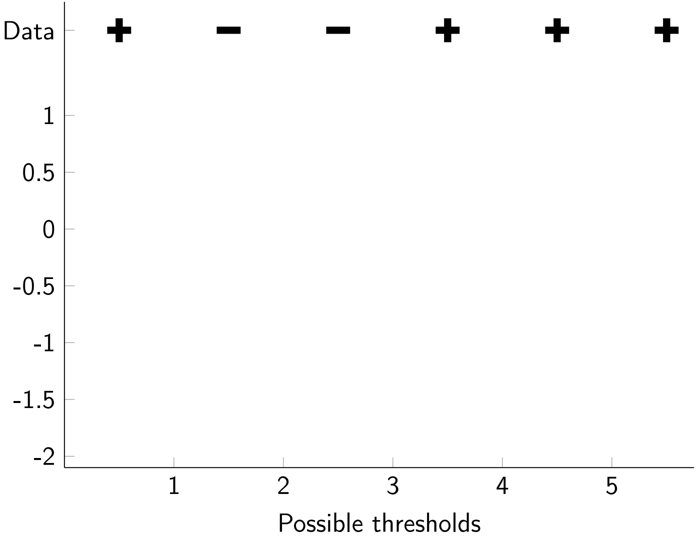
Likelihood-based splitting

If region R is split into A and B

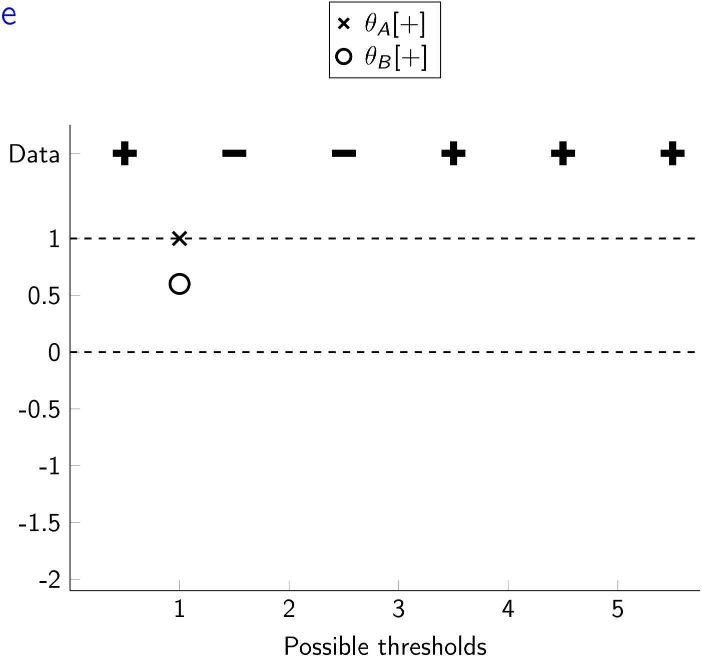
$$\begin{aligned}\triangle\mathcal{L}_{XY} &= \sum_{k=1}^{c}\left(n_A^{[k]}\log\theta_A[k] + n_B^{[k]}\log\theta_B[k] - n_R^{[k]}\log\theta_R[k]\right)\\ &= n_R H\left(\theta_R\right) - n_A H\left(\theta_A\right) - n_B H\left(\theta_B\right)\end{aligned}$$

$$H\left(\theta\right) := -\sum_{k=1}^{c}\theta[k]\log\theta[k]$$

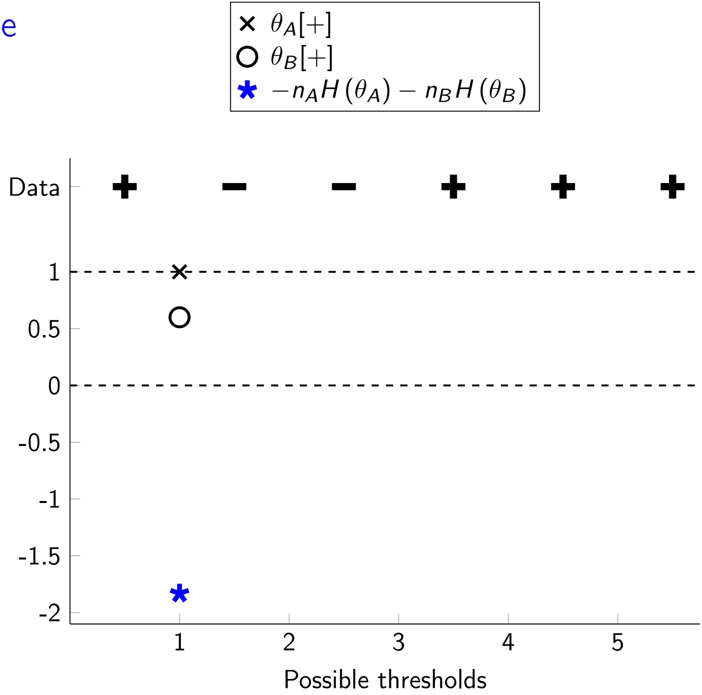
Example



Example

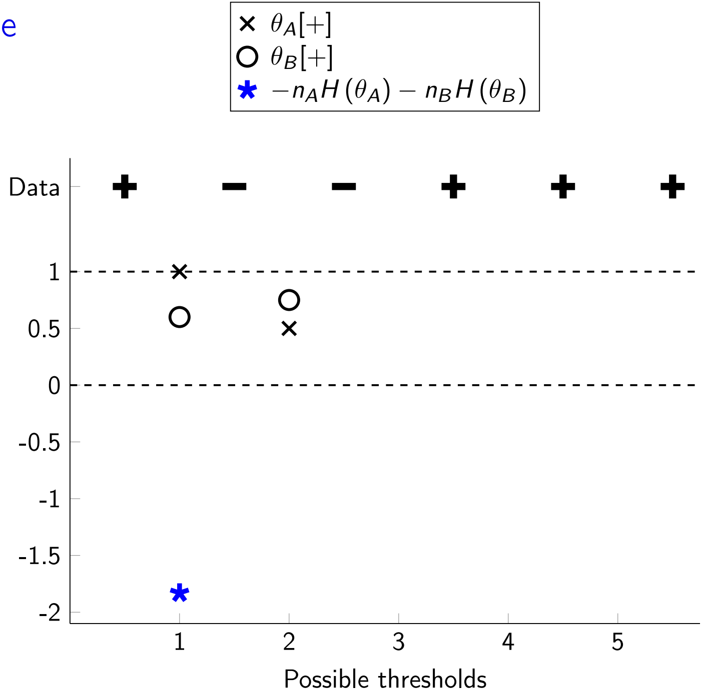


Example



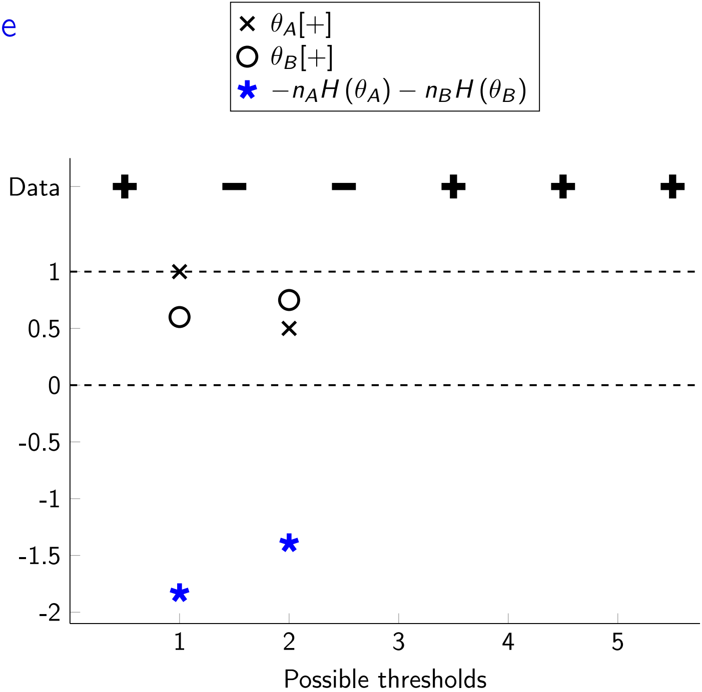
$\theta_A[+]$
$\theta_B[+]$
$-n_A H(\theta_A) - n_B H(\theta_B)$
Data
1
0.5
0
-0.5
-1
-1.5
-2
1
2
3
4
5
Possible thresholds

Example



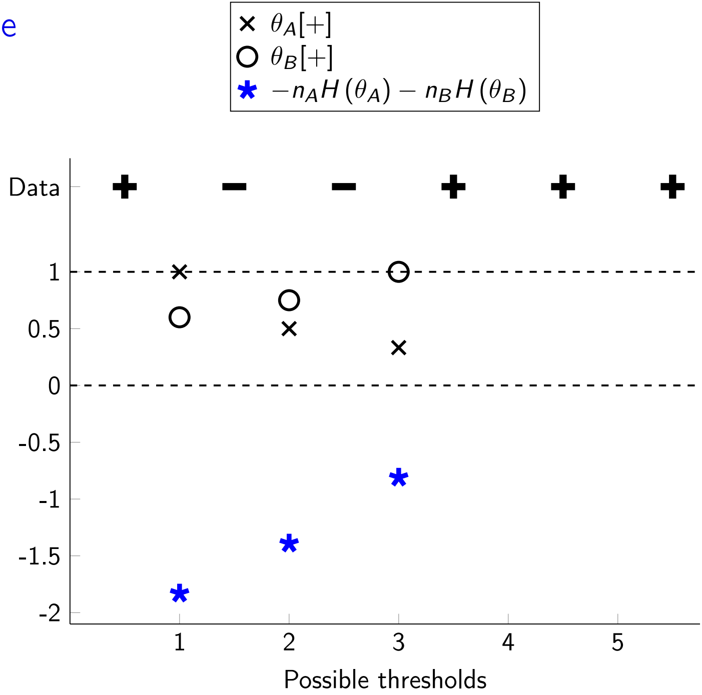
$\theta_A[+]$
$\theta_B[+]$
$-n_A H(\theta_A) - n_B H(\theta_B)$
Data
1
0.5
0
-0.5
-1
-1.5
-2
1
2
3
4
5
Possible thresholds

Example



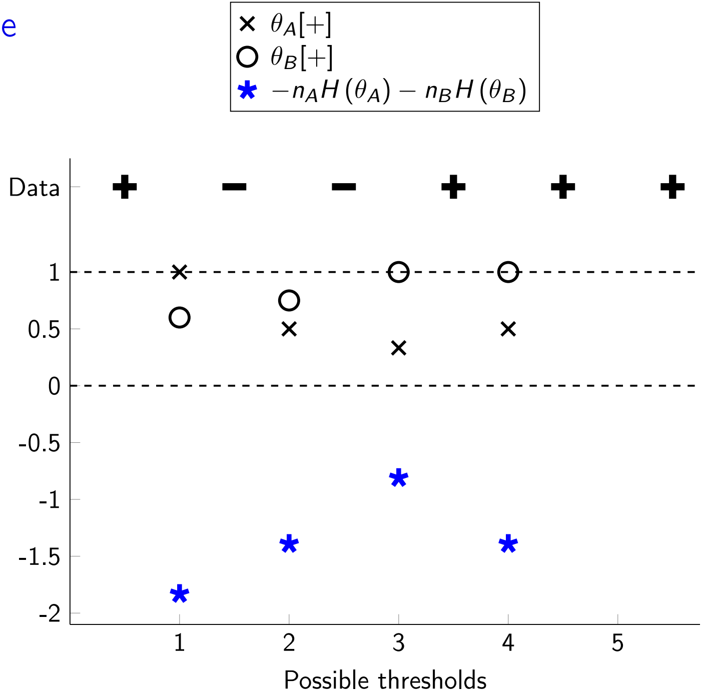
$\theta_A[+]$
$\theta_B[+]$
$-n_A H(\theta_A) - n_B H(\theta_B)$
Data
+
−
−
+
+
+
1
0.5
0
-0.5
-1
-1.5
-2
1
2
3
4
5
Possible thresholds

Example



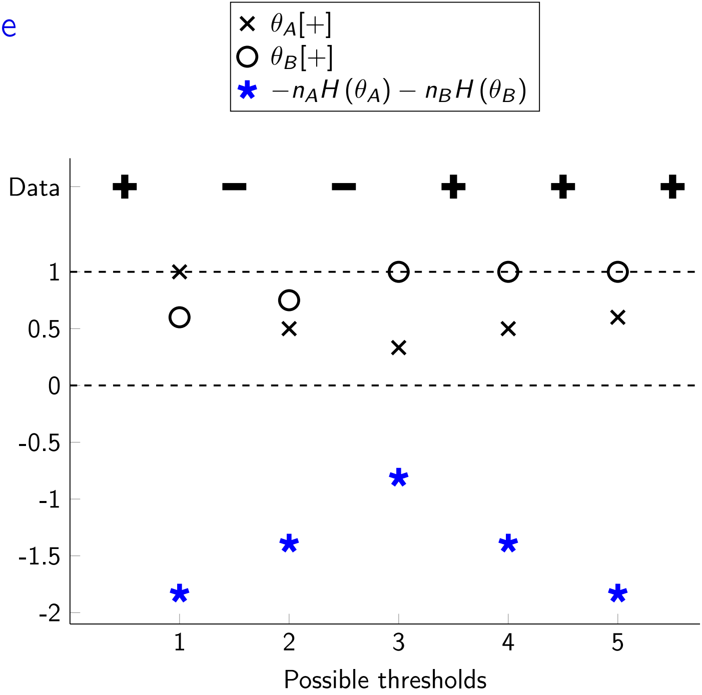
$\theta_A[+]$
$\theta_B[+]$
$-n_A H(\theta_A) - n_B H(\theta_B)$
Data
1
0.5
0
-0.5
-1
-1.5
-2
1
2
3
4
5
Possible thresholds

Example



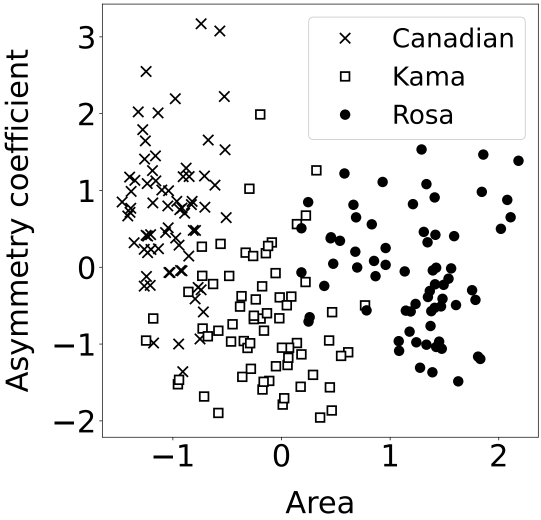
$\theta_A[+]$
$\theta_B[+]$
$-n_A H(\theta_A) - n_B H(\theta_B)$
Data
+
−
−
+
+
+
1
0.5
0
-0.5
-1
-1.5
-2
1
2
3
4
5
Possible thresholds

Example



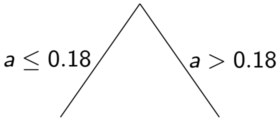
$\theta_A[+]$
$\theta_B[+]$
$-n_A H(\theta_A) - n_B H(\theta_B)$
Data
1
0.5
0
-0.5
-1
-1.5
-2
1
2
3
4
5
Possible thresholds

Wheat varieties



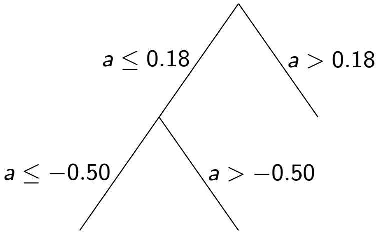
Recursive binary splitting

Features: area (a), asymmetric coefficient (c)



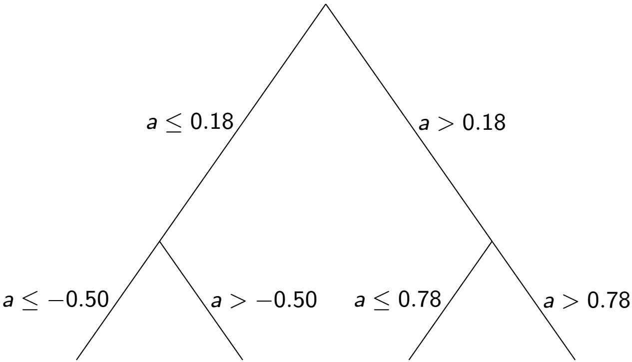
Recursive binary splitting

Features: area (a), asymmetric coefficient (c)



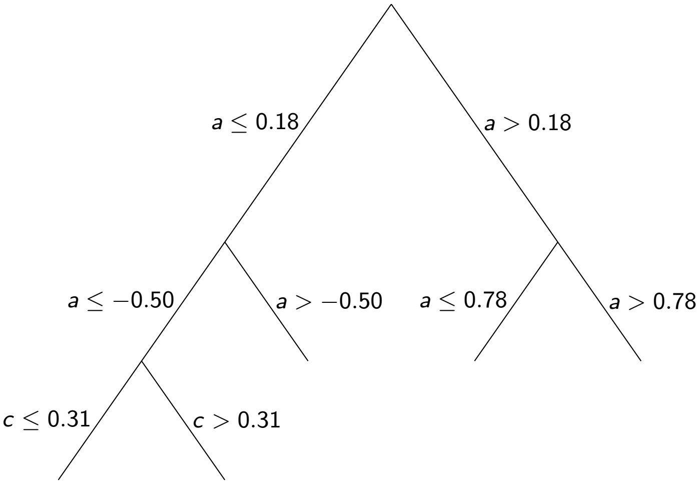
Recursive binary splitting

Features: area (a), asymmetric coefficient (c)



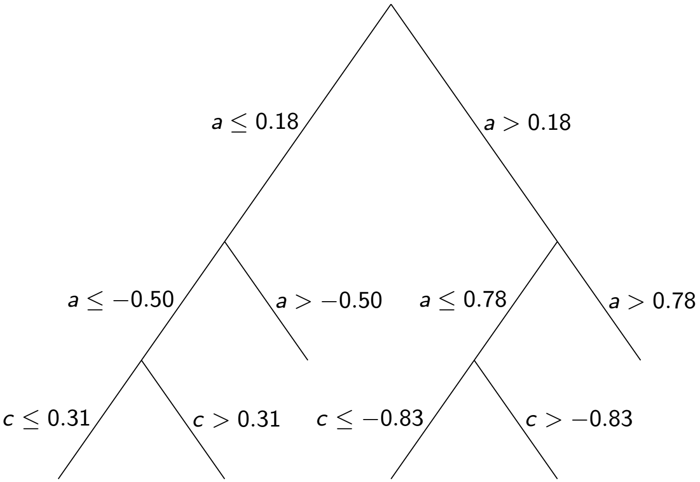
Recursive binary splitting

Features: area (a), asymmetric coefficient (c)

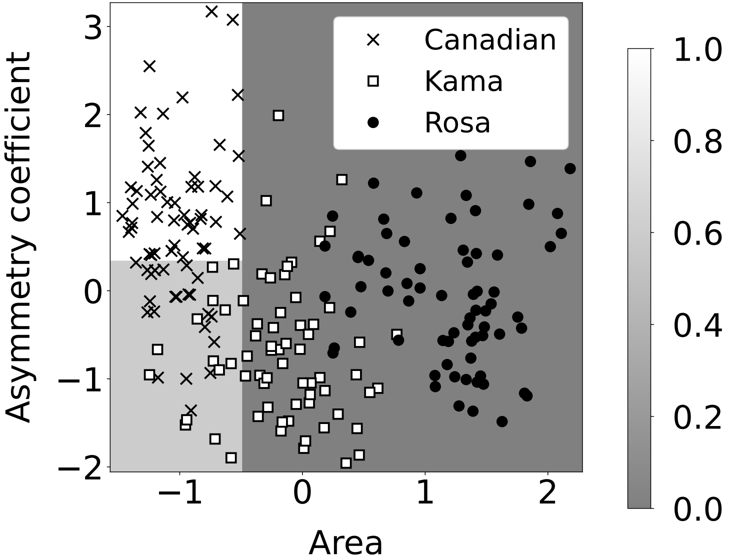


Recursive binary splitting

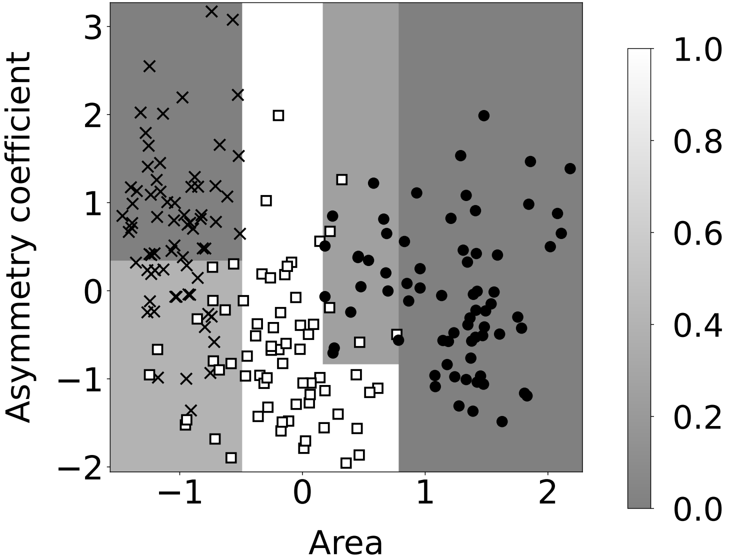
Features: area (a), asymmetric coefficient (c)



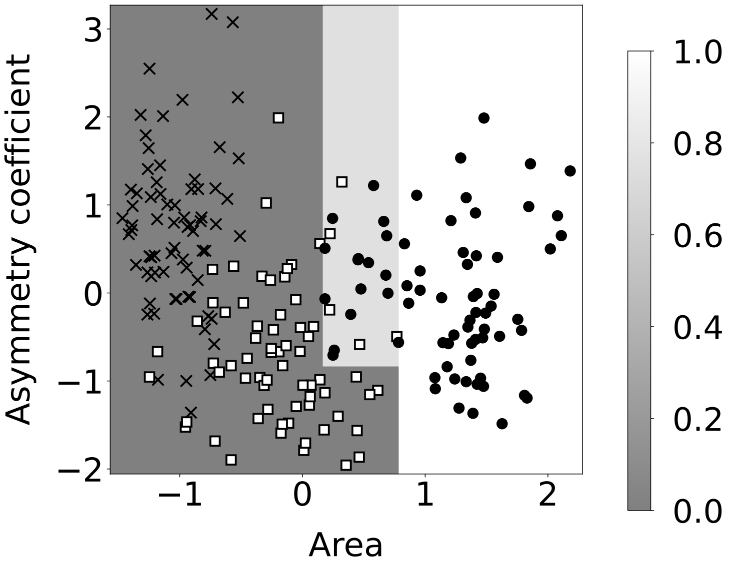
Canadian: Estimated probability



Kama: Estimated probability

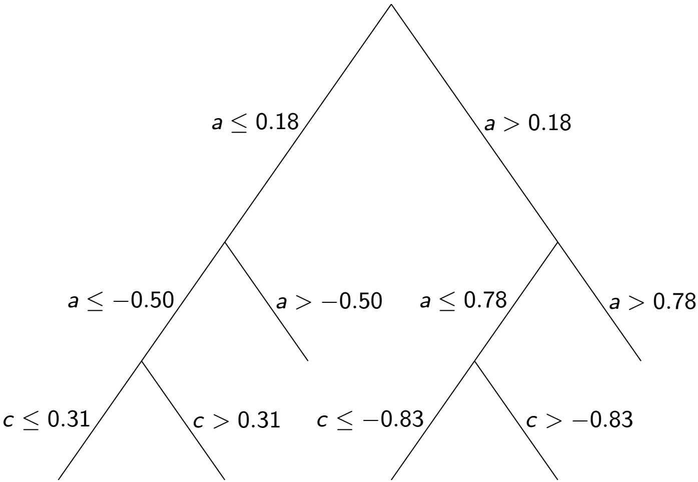


Rosa: Estimated probability



Interpretable!

Features: area (a), asymmetric coefficient (c)



What have we learned?

How to build nonlinear classification models using trees