CS145 Discussion

Week 3

Junheng, Shengming, Yunsheng

10/19/2018

Roadmap

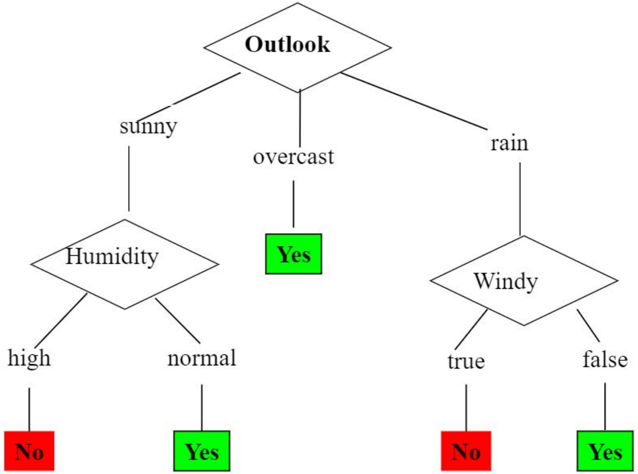
- Announcements
  - HW1 due Oct 19, 2018 (Friday, tonight)
  - Package your Report AND codes,README together and submit it through CCLE
- Review:
  - Decision Tree
    - Information Gain
    - Gain Ratio
    - Gini Index
  - SVM
    - Linear SVM
    - Soft Margin SVM
    - Non-linear SVM

Decision Tree

- Decision Tree Classification
  - Example: Play or Not?

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

-->



Decision Tree

- Choosing the Splitting Attribute
- At each node, available attributes are evaluated on the basis of separating the classes of the training examples.
- A Goodness function is used for this purpose:
    - Information Gain
    - Gain Ratio
    - Gini Index

A criterion for attribute selection

- Which is the best attribute?
  - The one which will result in the smallest tree
  - Heuristic: choose the attribute that produces the "purest" nodes
- Popular *impurity criterion*: *information gain*
  - Information gain increases with the average purity of the subsets that an attribute produces
- Strategy: choose attribute that results in greatest information gain

UCLA Entropy of a split

- Information in a split with ***x*** items of one class, ***y*** items of the second class

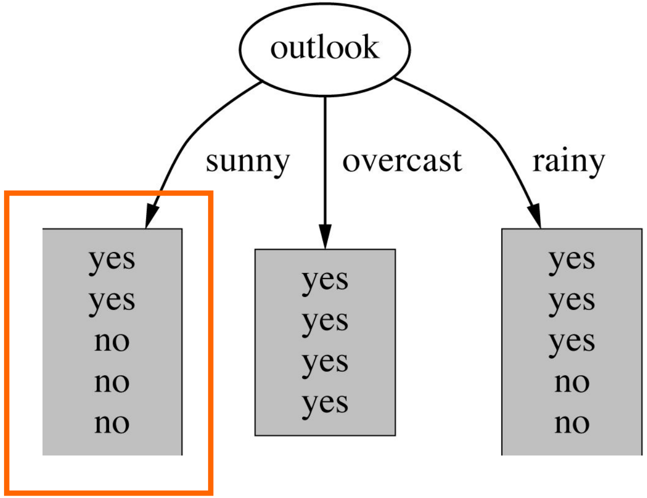
$$\text{info}([x, y]) = \text{entropy}\left(\frac{x}{x+y}, \frac{y}{x+y}\right)$$

$$= -\frac{x}{x+y}\log\left(\frac{x}{x+y}\right) - \frac{y}{x+y}\log\left(\frac{y}{x+y}\right)$$

Example: attribute “Outlook”

- “Outlook” = “Sunny”: 2 and 3 split

$$\text{info}([2,3]) = \text{entropy}(2/5,3/5) = -\frac{2}{5}\log(\frac{2}{5}) - \frac{3}{5}\log(\frac{3}{5}) = 0.971\ \text{bits}$$

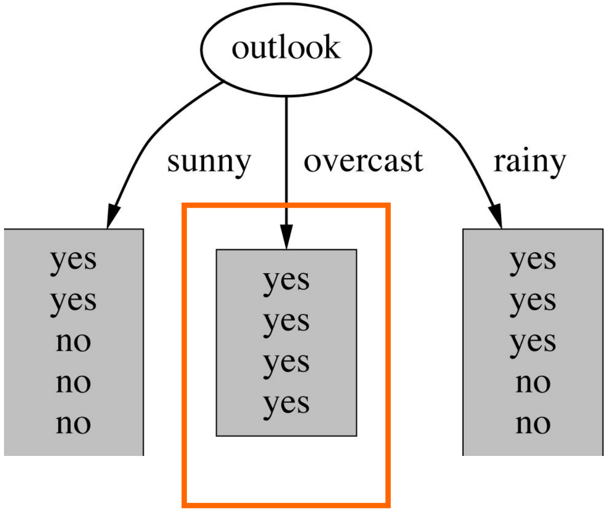


UCLA Outlook = Overcast

- "Outlook" = "Overcast": 4/0 split

$$\text{info}([4,0]) = \text{entropy}(1,0) = -1\log(1) - 0\log(0) = 0 \text{ bits}$$

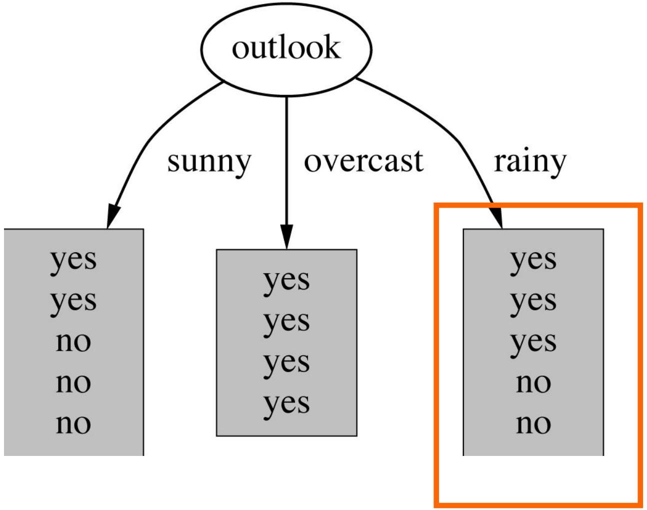
Note: log(0) is not defined, but we evaluate 0*log(0) as zero



Outlook = Rainy

- "Outlook" = "Rainy":

$$\text{info}([3,2]) = \text{entropy}(3/5, 2/5) = -\frac{3}{5}\log(\frac{3}{5}) - \frac{2}{5}\log(\frac{2}{5}) = 0.971 \text{ bits}$$



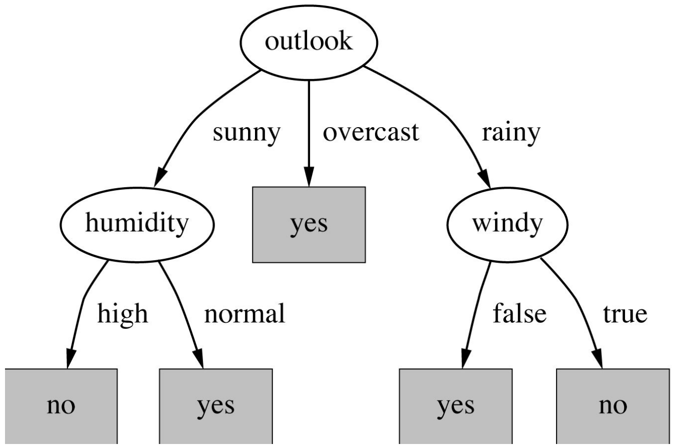
UCLA Expected Information

Expected information for attribute:

$$\text{info}([3,2],[4,0],[3,2]) = (5/14)\times 0.971 + (4/14)\times 0 + (5/14)\times 0.971$$

$$= 0.693 \text{ bits}$$

The final decision tree



- Note: not all leaves need to be pure; sometimes identical instances have different classes
  - ⇒ Splitting stops when data can't be split any further

UCLA Computing the information gain

- Information gain:

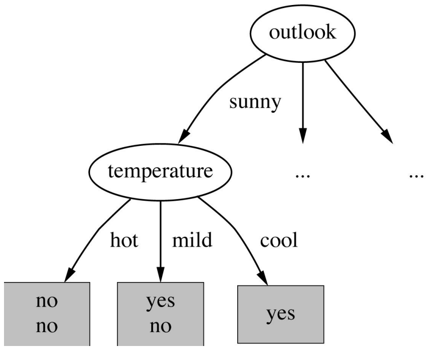
(information before split) – (information after split)

$$\text{gain}(\text{"Outlook"}) = \text{info}([9,5]) - \text{info}([2,3],[4,0],[3,2]) = 0.940 - 0.693$$
$$= 0.247 \text{ bits}$$

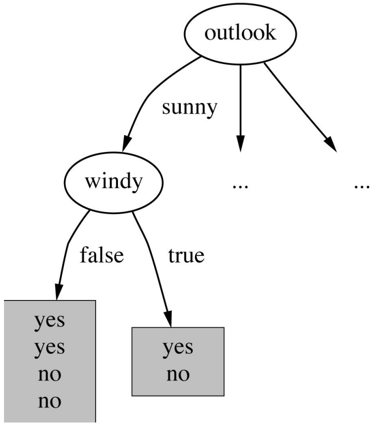
- Information gain for attributes from weather data:

$$\text{gain}(\text{"Outlook"}) = 0.247 \text{ bits}$$
$$\text{gain}(\text{"Temperature"}) = 0.029 \text{ bits}$$
$$\text{gain}(\text{"Humidity"}) = 0.152 \text{ bits}$$
$$\text{gain}(\text{"Windy"}) = 0.048 \text{ bits}$$

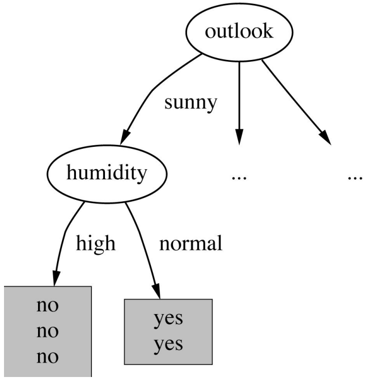
UCLA Continuing to split



gain("Temperature") = 0.571 bits

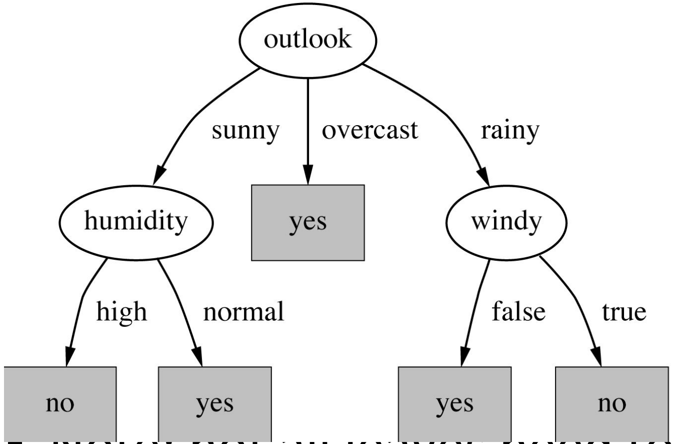


gain("Windy") = 0.020 bits



gain("Humidity") = 0.971 bits

UCLA The final decision tree



- Note: not all leaves need to be pure; sometimes identical instances have different classes

  ⇒ Splitting stops when data can't be split any further

Decision Tree

Gain Ratio

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$

Gain Ratio = Gain_A(D) / SplitInfo_A(D)

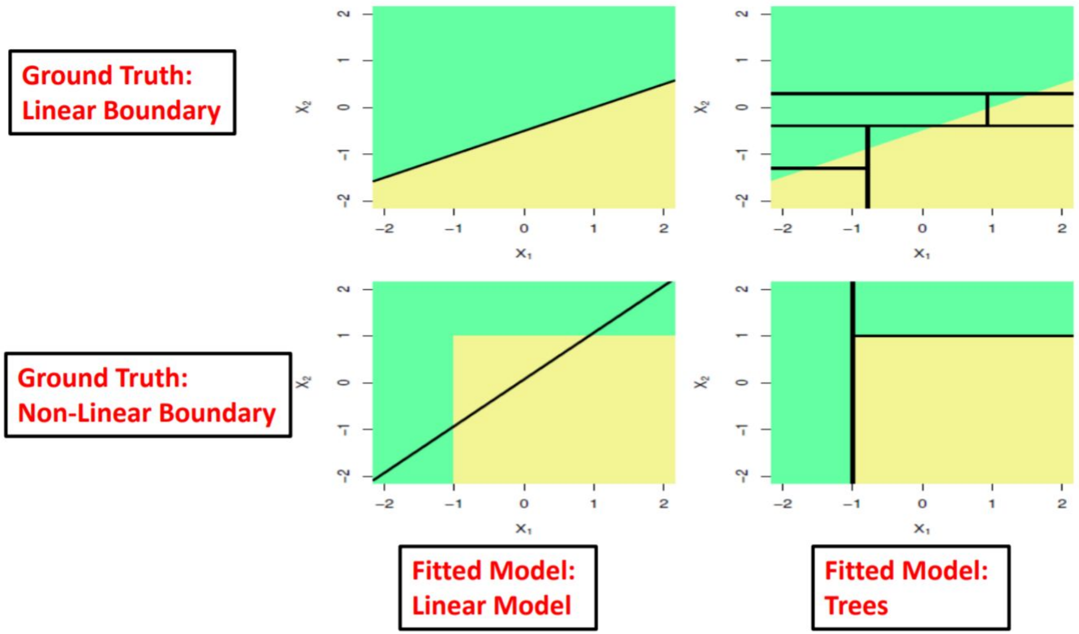
Why Gain Ratio?

Unbiased compared with Information Gain

Why? (https://stats.stackexchange.com/questions/306456/how-is-information-gain-biased)

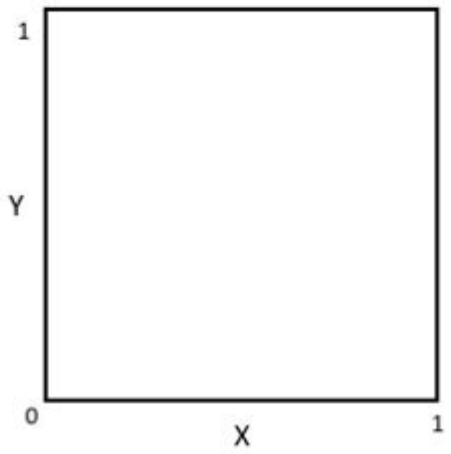
Decision Tree

- Is the decision boundary for decision tree linear?   No



Visual Tutorials of Decision Trees

https://algobeans.com/2016/07/27/decision-trees-tutorial/



UCLA

Support Vector Machine

Hyperplane separating the data points

$$\boldsymbol{w}^{\boldsymbol{T}}\mathbf{x} + \mathrm{b} = 0$$

Maximize margin

$$\rho = \frac{2}{||w||}$$

Solution

$$\boldsymbol{w} = \sum \alpha_i y_i \mathbf{x}_i \qquad b = \sum\nolimits_{k:\alpha_k \neq 0} (y_k - \boldsymbol{w}^{\boldsymbol{T}}\mathbf{x}_k)/N_k$$

Margin Formula

Margin Lines

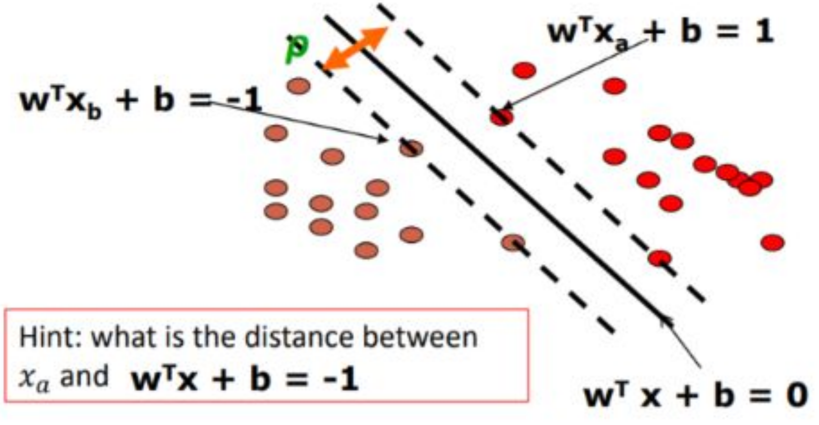
$$\boldsymbol{w}^T\mathbf{x}_a + \mathrm{b} = 1 \qquad \boldsymbol{w}^T\mathbf{x}_b + \mathrm{b} = -1$$

Distance between parallel lines

$$d = \frac{|c_2 - c_1|}{\sqrt{a^2 + b^2}}$$

Margin

$$\rho = \frac{|(b+1) - (b-1)|}{||w||} = \frac{2}{||w||}$$



Linear SVM Example

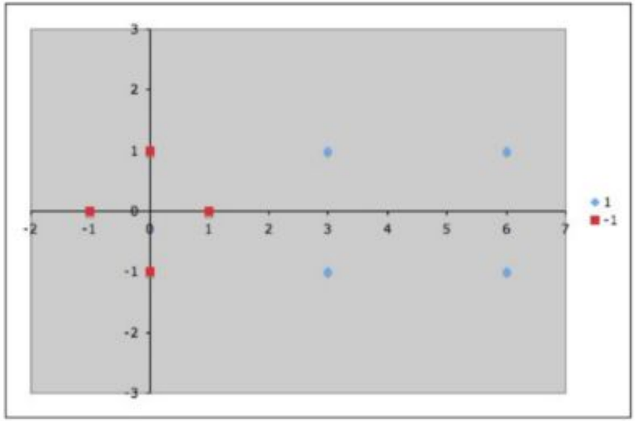
- Positively labeled data points (1 to 4)

$$\left\{\begin{pmatrix}3\\1\end{pmatrix},\begin{pmatrix}3\\-1\end{pmatrix},\begin{pmatrix}6\\1\end{pmatrix},\begin{pmatrix}6\\-1\end{pmatrix}\right\}$$

- Negatively labeled data points (5 to 8)

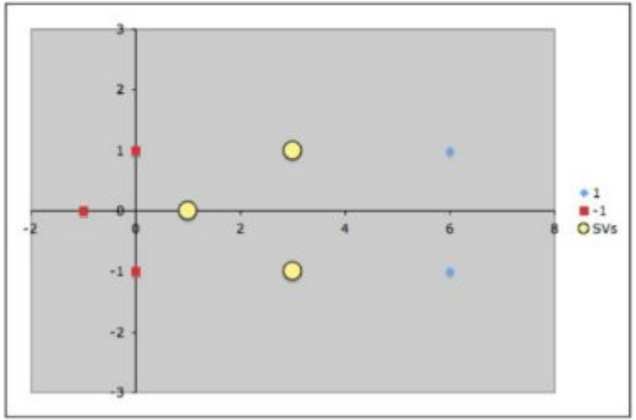
$$\left\{\begin{pmatrix}1\\0\end{pmatrix},\begin{pmatrix}0\\1\end{pmatrix},\begin{pmatrix}0\\-1\end{pmatrix},\begin{pmatrix}-1\\0\end{pmatrix}\right\}$$

- Alpha values
  - $\alpha_1 = 0.75$
  - $\alpha_2 = 0.75$
  - $\alpha_5 = 3.5$
  - Others = 0



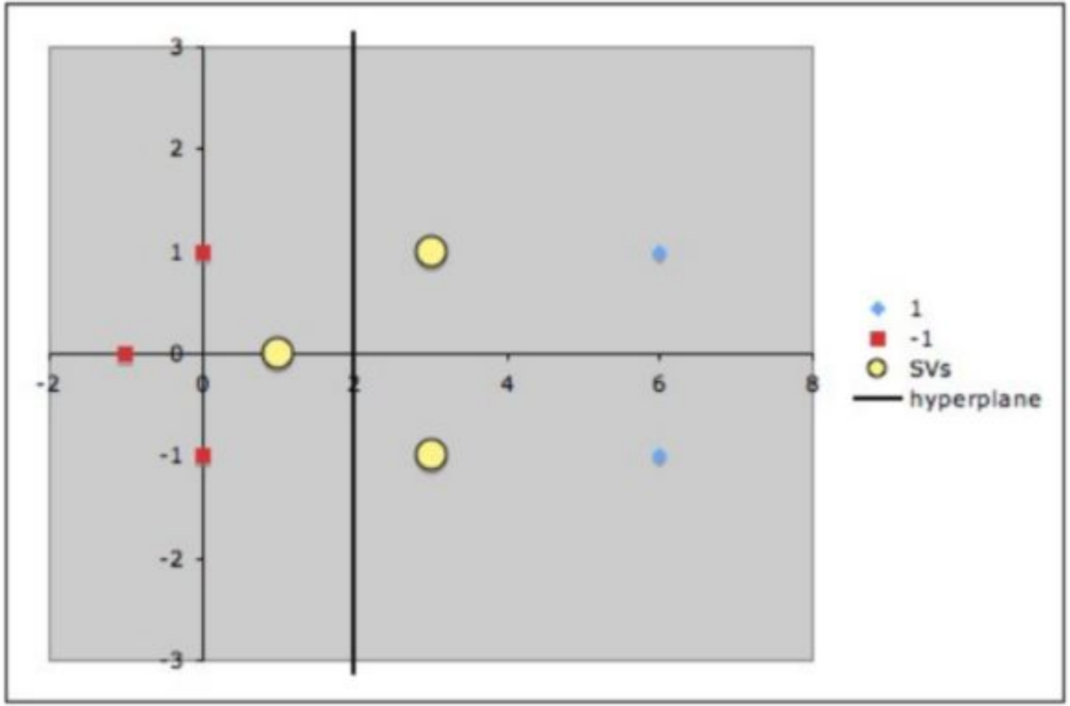
Linear SVM Example

- Which points are support vectors?
- Calculate normal vector of hyperplane: $\boldsymbol{w}$
- Calculate the bias term
- What is the decision boundary?
- Predict class of new point (4, 1)



$$\mathbf{w} = \sum \alpha_i y_i \mathbf{x}_i \qquad b = \sum_{k:\alpha_k \neq 0} (y_k - \mathbf{w}^T \mathbf{x}_k)/N_k$$

Plot



Non-linear SVM Example

- Positively labeled data points (1 to 4)

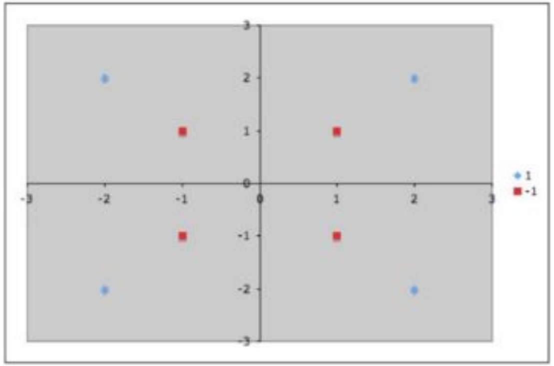
$$\left\{\begin{pmatrix}2\\2\end{pmatrix},\begin{pmatrix}2\\-2\end{pmatrix},\begin{pmatrix}-2\\-2\end{pmatrix},\begin{pmatrix}-2\\2\end{pmatrix}\right\}$$

- Negatively labeled data points (5 to 8)

$$\left\{\begin{pmatrix}1\\1\end{pmatrix},\begin{pmatrix}1\\-1\end{pmatrix},\begin{pmatrix}-1\\-1\end{pmatrix},\begin{pmatrix}-1\\1\end{pmatrix}\right\}$$

- Non-linear mapping

$$\Phi_1\begin{pmatrix}x_1\\x_2\end{pmatrix} = \begin{cases}\begin{pmatrix}4-x_2\\4-x_1\end{pmatrix} & \text{if } \sqrt{x_1^2+x_2^2} > 2\\ \begin{pmatrix}x_1\\x_2\end{pmatrix} & \text{otherwise}\end{cases}$$



Non-linear SVM Example

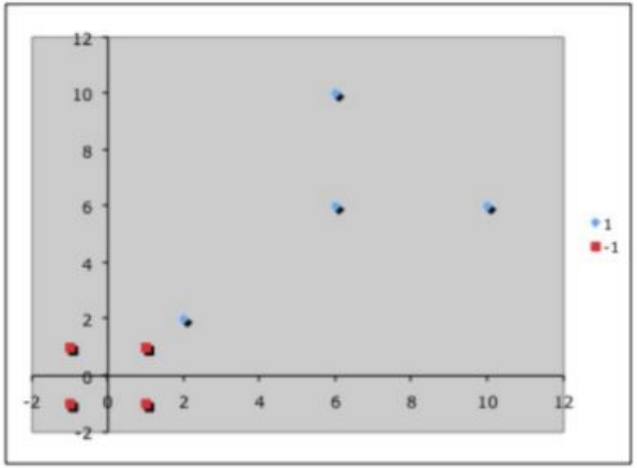
- New positively labeled data points (1 to 4)

$$\left\{\begin{pmatrix}2\\2\end{pmatrix},\begin{pmatrix}6\\2\end{pmatrix},\begin{pmatrix}6\\6\end{pmatrix},\begin{pmatrix}2\\6\end{pmatrix}\right\}$$

- New negatively labeled data points (5 to 8)

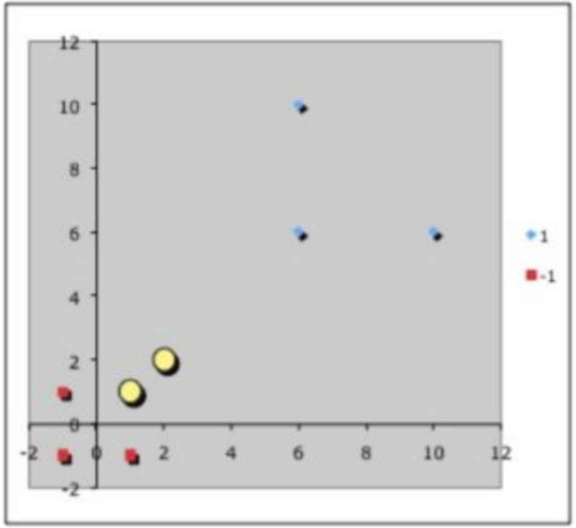
$$\left\{\begin{pmatrix}1\\1\end{pmatrix},\begin{pmatrix}1\\-1\end{pmatrix},\begin{pmatrix}-1\\-1\end{pmatrix},\begin{pmatrix}-1\\1\end{pmatrix}\right\}$$

- Alpha values
  - $\alpha_1 = 4$
  - $\alpha_5 = 7$
  - Others = 0



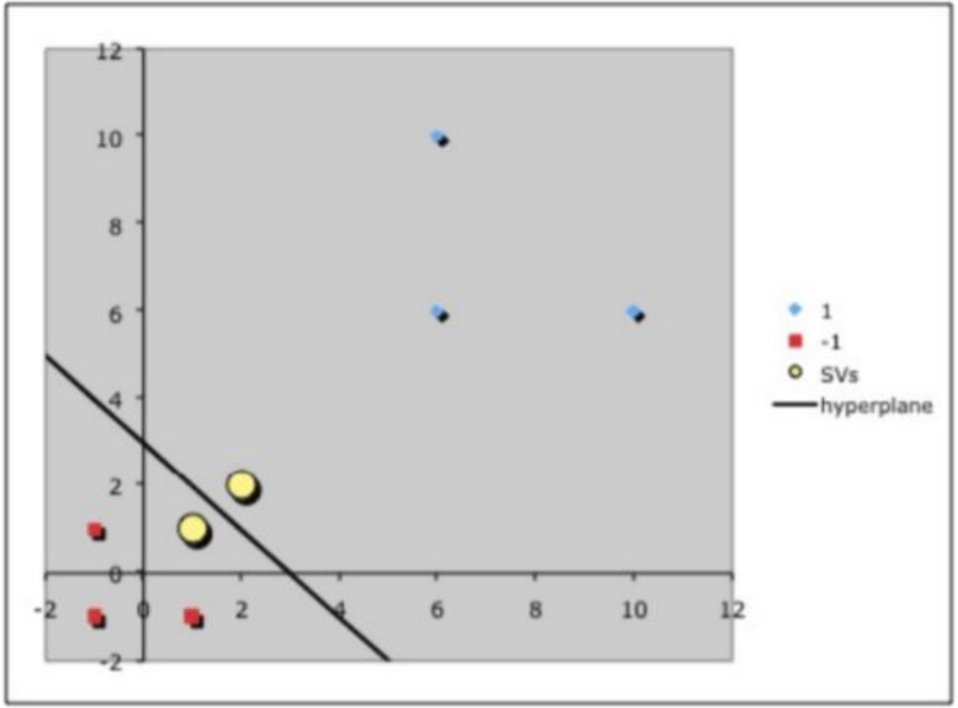
Non-linear SVM Example

- Which points are support vectors?
- Calculate normal vector of hyperplane: $\boldsymbol{w}$
- Calculate the bias term
- What is the decision boundary?
- Predict class of new point (4, 5)



UCLA

Plot



Visualize Tutorials of Decision Trees

http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

http://explained.ai/decision-tree-viz/

Visual Tutorials of SVM

https://cs.stanford.edu/people/karpathy/svmjs/demo/

Thank you!

Q & A

Blank

Double-row title

Subtitle

Single-row title

Figure Page