# JSC 270 - Lecture 5

# Classifiers

https://jsc270.github.io/

# Announcements

- Guest lecture by Benjamin Haibe-Kains on reproducibility today 2-3pm
- Perusall assignment is online, due on Saturday night
- Presentation for Assignment 2 is due on Feb 11th (instructions are on Quercus)
- Assignment 3 will be up online on Feb 12th
- Survey to give us feedback about how we are doing
  https://forms.gle/wYGYuSaCgNDbQgBv8

# What classifiers have you heard of?

# Data

N - number of records (samples, entries)

M - number of features (predictors)

D is N x M

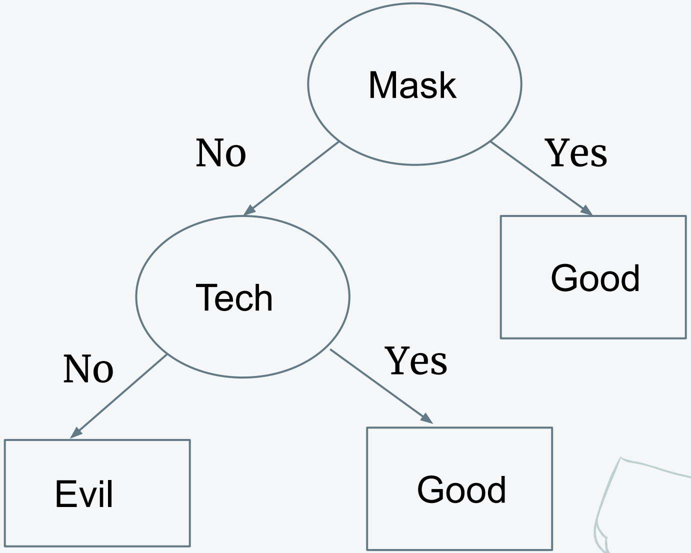
N = ?

M = ?

| Hero Name | Mask | Cape | Tech | Pointy Ears | Height (cm) | Good/ evil |
|---|---|---|---|---|---|---|
| **Ant-Man** | False | False | True | False | 183 | Good |
| **Elastagirl** | True | False | False | False | 172 | Good |
| **Batman** | True | True | True | True | 188 | Good |
| **Robin** | True | True | True | True | 165 | Good |
| **Catwoman** | True | True | False | True | 175 | Evil |
| **Penguin** | False | False | False | False | 147 | Evil |
| **Dormammu** | False | False | False | False | 185 | Evil |
| **Joker** | False | False | False | False | 184 | Evil |

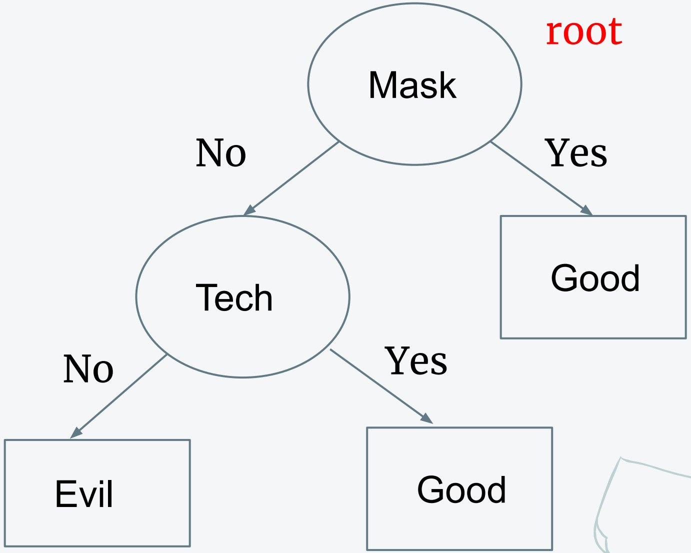
# Decision Trees Example.

| Hero Name | Mask | Cape | Tech | Pointy Ears | Height (cm) | Good/ evil |
|---|---|---|---|---|---|---|
| **Ant-Man** | False | False | True | False | 183 | Good |
| **Elastagirl** | True | False | False | False | 172 | Good |
| **Batman** | True | True | True | True | 188 | Good |
| **Robin** | True | True | True | True | 165 | Good |
| **Catwoman** | True | True | False | True | 175 | Evil |
| **Penguin** | False | False | True | False | 147 | Evil |
| **Dormammu** | False | False | False | False | 185 | Evil |
| **Joker** | False | False | False | False | 184 | Evil |



# Decision Trees Example. Definitions

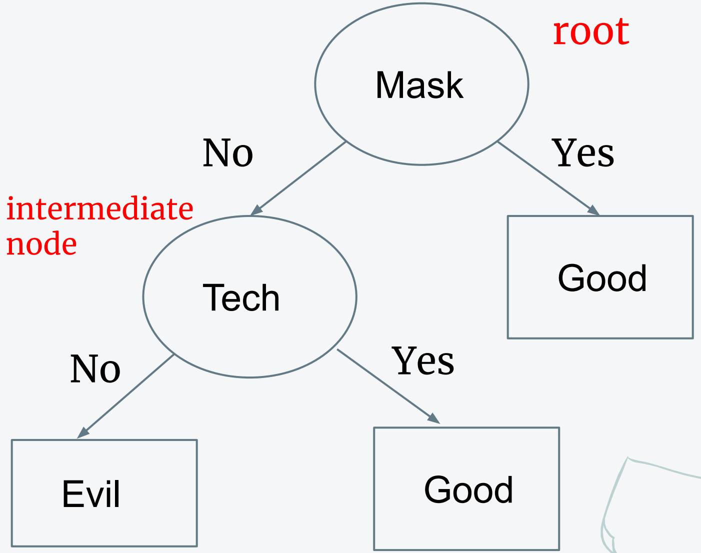
| Hero Name | Mask | Cape | Tech | Pointy Ears | Height (cm) | Good/ evil |
|---|---|---|---|---|---|---|
| Ant-Man | False | False | True | False | 183 | Good |
| Elastagirl | True | False | False | False | 172 | Good |
| Batman | True | True | True | True | 188 | Good |
| Robin | True | True | True | True | 165 | Good |
| Catwoman | True | True | False | True | 175 | Evil |
| Penguin | False | False | True | False | 147 | Evil |
| Dormammu | False | False | False | False | 185 | Evil |
| Joker | False | False | False | False | 184 | Evil |




Depth = 2 (length of the longest path)

# Decision Trees Example. Definitions

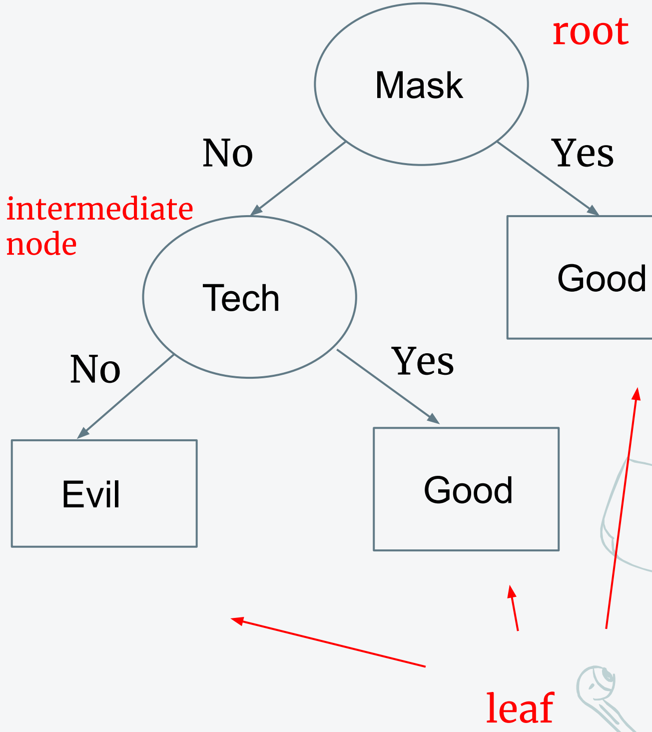
| Hero Name | Mask | Cape | Tech | Pointy Ears | Height (cm) | Good/ evil |
|---|---|---|---|---|---|---|
| **Ant-Man** | False | False | True | False | 183 | Good |
| **Elastagirl** | True | False | False | False | 172 | Good |
| **Batman** | True | True | True | True | 188 | Good |
| **Robin** | True | True | True | True | 165 | Good |
| **Catwoman** | True | True | False | True | 175 | Evil |
| **Penguin** | False | False | True | False | 147 | Evil |
| **Dormammu** | False | False | False | False | 185 | Evil |
| **Joker** | False | False | False | False | 184 | Evil |




Depth = 2 (length of the longest path)

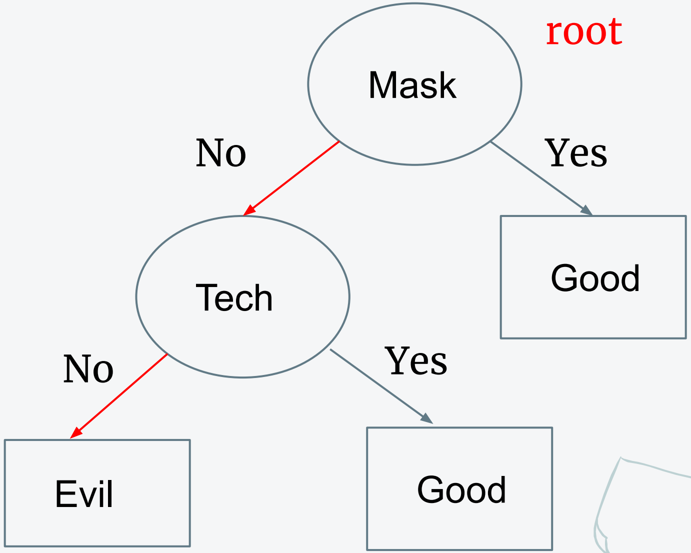
# Decision Trees Example. Definitions

| Hero Name | Mask | Cape | Tech | Pointy Ears | Height (cm) | Good/ evil |
|---|---|---|---|---|---|---|
| **Ant-Man** | False | False | True | False | 183 | Good |
| **Elastagirl** | True | False | False | False | 172 | Good |
| **Batman** | True | True | True | True | 188 | Good |
| **Robin** | True | True | True | True | 165 | Good |
| **Catwoman** | True | True | False | True | 175 | Evil |
| **Penguin** | False | False | True | False | 147 | Evil |
| **Dormammu** | False | False | False | False | 185 | Evil |
| **Joker** | False | False | False | False | 184 | Evil |



# Decision Trees Example. Definitions

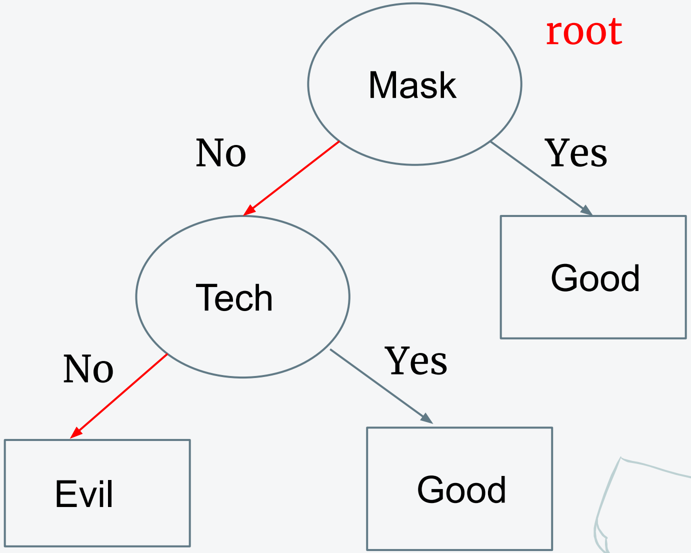
| Hero Name | Mask | Cape | Tech | Pointy Ears | Height (cm) | Good/ evil |
|---|---|---|---|---|---|---|
| **Ant-Man** | False | False | True | False | 183 | Good |
| **Elastagirl** | True | False | False | False | 172 | Good |
| **Batman** | True | True | True | True | 188 | Good |
| **Robin** | True | True | True | True | 165 | Good |
| **Catwoman** | True | True | False | True | 175 | Evil |
| **Penguin** | False | False | True | False | 147 | Evil |
| **Dormammu** | False | False | False | False | 185 | Evil |
| **Joker** | False | False | False | False | 184 | Evil |




Depth – length of the longest path
Depth = ?

# Decision Trees Example. Definitions

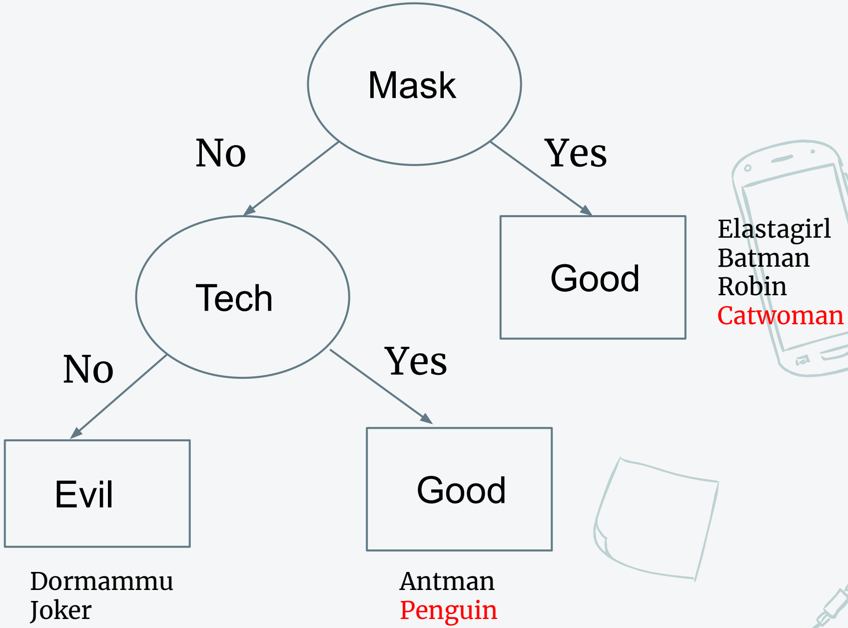
| Hero Name | Mask | Cape | Tech | Pointy Ears | Height (cm) | Good/ evil |
|---|---|---|---|---|---|---|
| Ant-Man | False | False | True | False | 183 | Good |
| Elastagirl | True | False | False | False | 172 | Good |
| Batman | True | True | True | True | 188 | Good |
| Robin | True | True | True | True | 165 | Good |
| Catwoman | True | True | False | True | 175 | Evil |
| Penguin | False | False | True | False | 147 | Evil |
| Dormammu | False | False | False | False | 185 | Evil |
| Joker | False | False | False | False | 184 | Evil |




Each path is a classification rule!
No mask, no tech -> evil

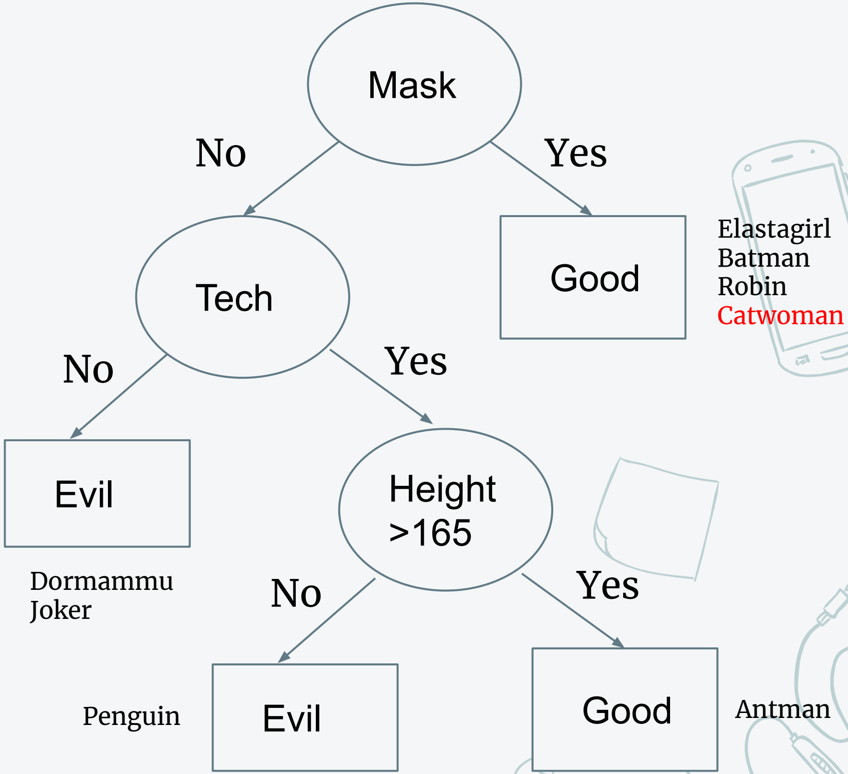
# Decision Trees Example. A possible tree.

| Hero Name | Mask | Cape | Tech | Pointy Ears | Height (cm) | Good/ evil |
|---|---|---|---|---|---|---|
| **Ant-Man** | False | False | True | False | 183 | Good |
| **Elastagirl** | True | False | False | False | 172 | Good |
| **Batman** | True | True | True | True | 188 | Good |
| **Robin** | True | True | True | True | 165 | Good |
| **Catwoman** | True | True | False | True | 175 | Evil |
| **Penguin** | False | False | True | False | 147 | Evil |
| **Dormammu** | False | False | False | False | 185 | Evil |
| **Joker** | False | False | False | False | 184 | Evil |



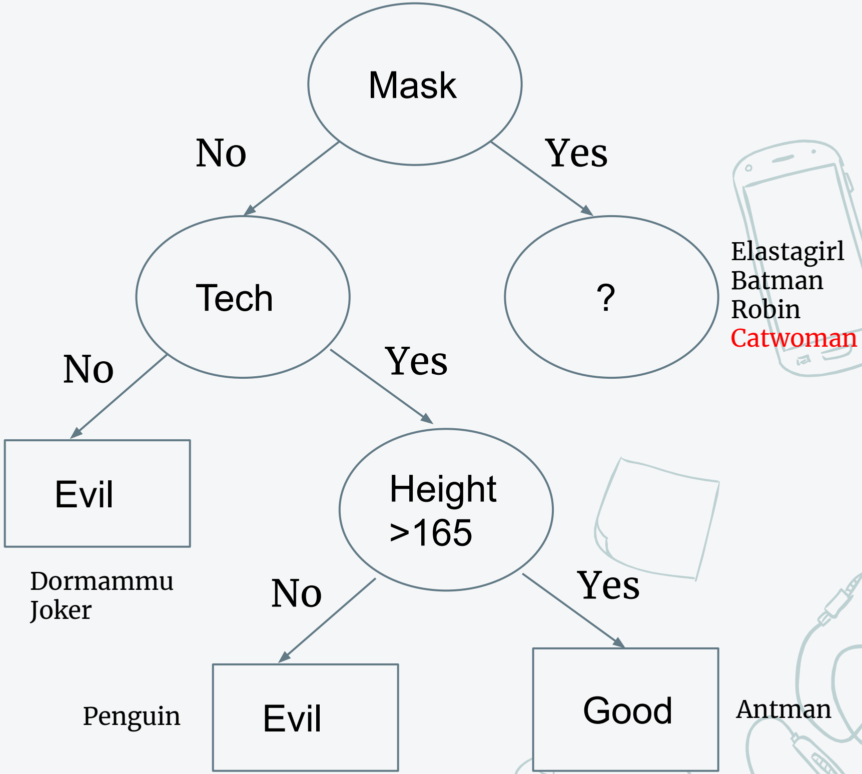
# Decision Trees Example. A possible Tree

| Hero Name | Mask | Cape | Tech | Pointy Ears | Height (cm) | Good/ evil |
|---|---|---|---|---|---|---|
| **Ant-Man** | False | False | True | False | 183 | Good |
| **Elastagirl** | True | False | False | False | 172 | Good |
| **Batman** | True | True | True | True | 188 | Good |
| **Robin** | True | True | True | True | 165 | Good |
| **Catwoman** | True | True | False | True | 175 | Evil |
| **Penguin** | False | False | True | False | 147 | Evil |
| **Dormammu** | False | False | False | False | 185 | Evil |
| **Joker** | False | False | False | False | 184 | Evil |



# Decision Trees Example. A possible Tree

| Hero Name | Mask | Cape | Tech | Pointy Ears | Height (cm) | Good/ evil |
|---|---|---|---|---|---|---|
| **Ant-Man** | False | False | True | False | 183 | Good |
| **Elastagirl** | True | False | False | False | 172 | Good |
| **Batman** | True | True | True | True | 188 | Good |
| **Robin** | True | True | True | True | 165 | Good |
| **Catwoman** | True | True | False | True | 175 | Evil |
| **Penguin** | False | False | True | False | 147 | Evil |
| **Dormammu** | False | False | False | False | 185 | Evil |
| **Joker** | False | False | False | False | 184 | Evil |



# How do you find a good tree?

Important concept: Impurity Function

Impurity? As it sounds – error in class assignment

Most common impurity functions:

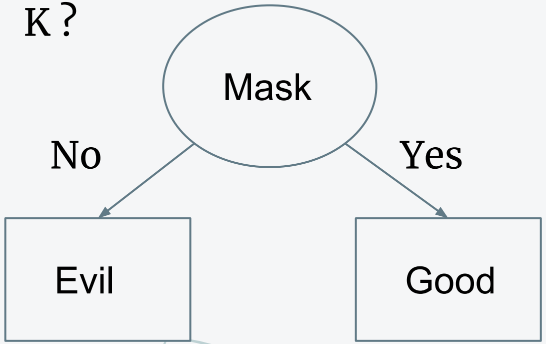
- Gini
- Entropy
- Gain Ratio

# Some annotation

S - data to split into K classes: $\{(\mathbf{x}_1,y_1), (\mathbf{x}_2,y_2),\ldots,(\mathbf{x}_n,y_n)\}$

$p_k = |S_k|/|S|$, where $S_k = \{(\mathbf{x}_j,y_j) \in S \text{ for all } j: y_j = k\}$

K ?



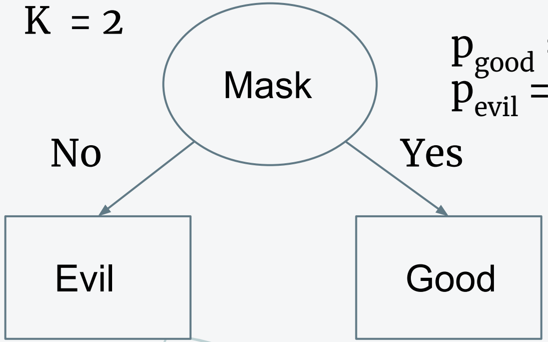
# Some annotation

S - data to split into K classes: $\{(\mathbf{x}_1,y_1), (\mathbf{x}_2,y_2),\ldots,(\mathbf{x}_n,y_n)\}$

$p_k = |S_k|/|S|$, where $S_k = \{(\mathbf{x}_j,y_j) \in S \text{ for all } j: y_j = k\}$

K = 2

$p_{good}$ =
$p_{evil}$ =




$p_{good}$ =
$p_{evil}$ =

$p_{good}$ =
$p_{evil}$ =

# What's the worst possible split?

# Gini Impurity

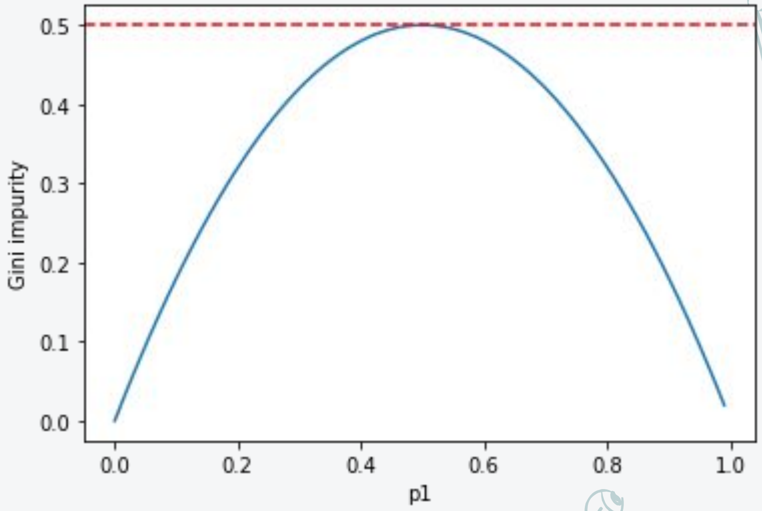
$$G(S) = \sum_{k=1}^{K} p_k(1 - p_k)$$

For K = 2: $p_1$, $p_2$, where $p_1+p_2=1$

Then $G(S) = p_1(1-p_1) + p_2(1-p_2)$

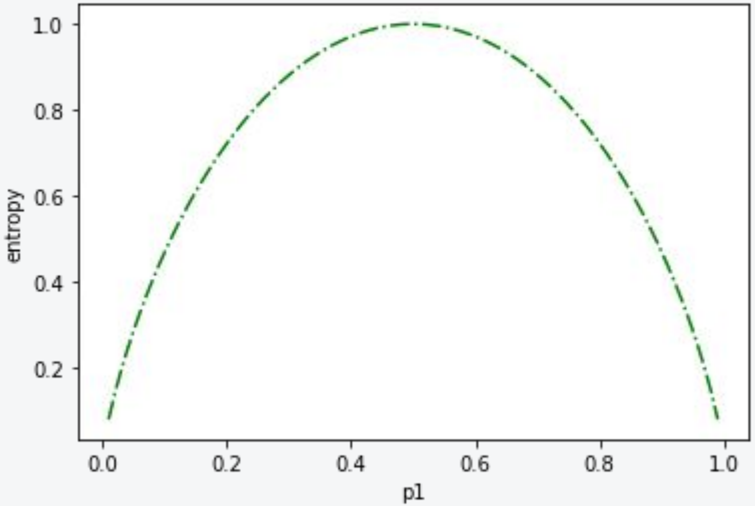
$= p_1p_2 + p_2p_1 = 2p_1p_2$

$= 2p_1(1-p_1)$

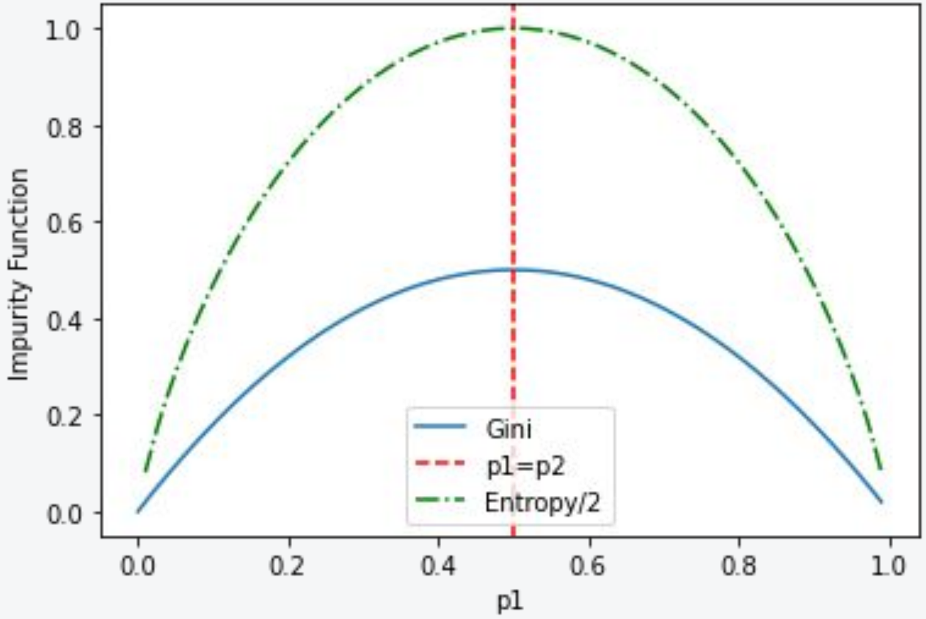


# Entropy

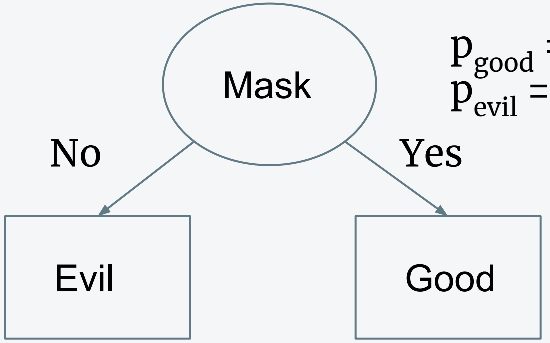
$$H(S) = -\sum_{k=1}^{K} p_k log(p_k)$$



# Comparison of G(S) and H(S)



# Computing Impurity




$p_{good} = 0.5$
$p_{evil} = 0.5$

Gini: $G(Mask) =$
Entropy: $H(Mask) =$

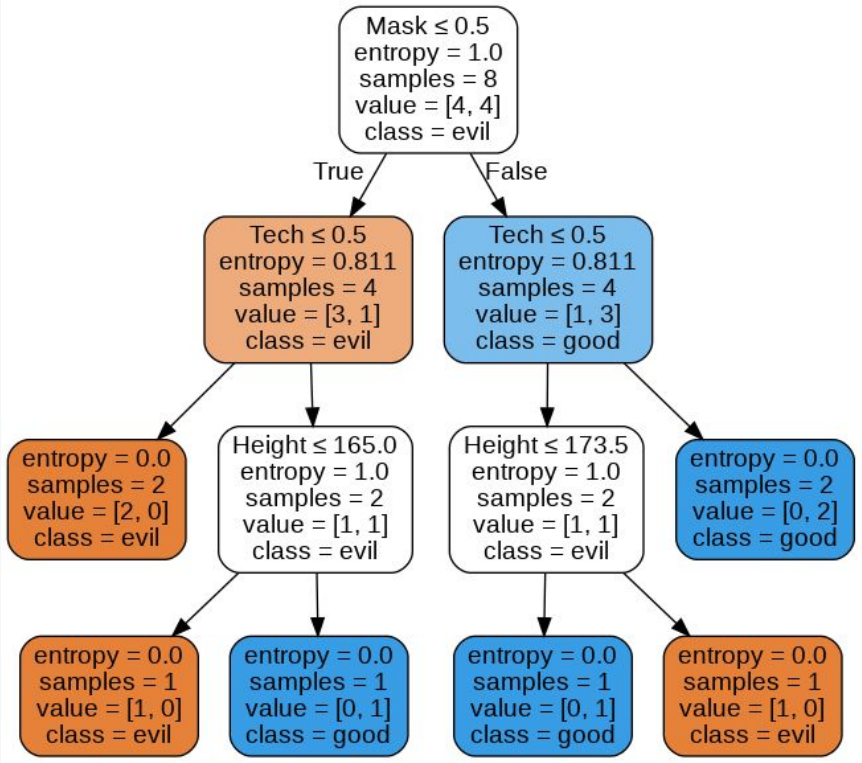
# How to construct a tree using an Impurity function?

NP-hard to find the best tree

Greedy approach:

1.
2.
3.

# A tree from our superhero example




Impurity = entropy
Depth = ?

Any problems?

DecisionTreeClassifier from the scikitlearn tree package

# What's overfitting?

Definition:

Is it high bias or high variance scenario?

# Decision Trees Summary

Advantages?

Disadvantages?

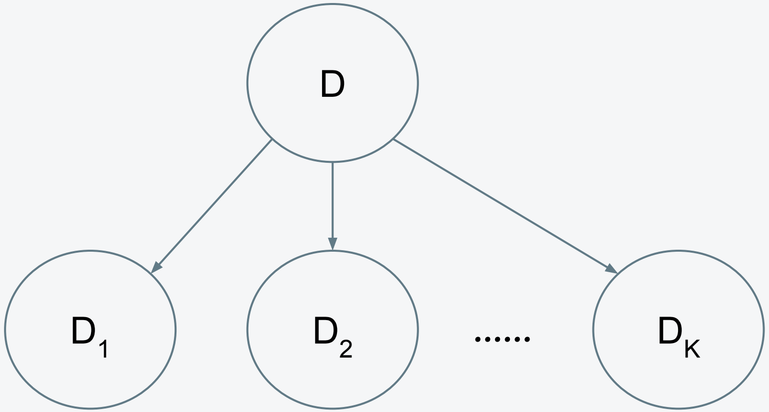
# Bagging

An effective way to reduce the variance of a classifier

Helps with high variance classifiers (overfitting)

# Bagging

N - number of records (samples, entries)
M - number of features (predictors)
D is N x M

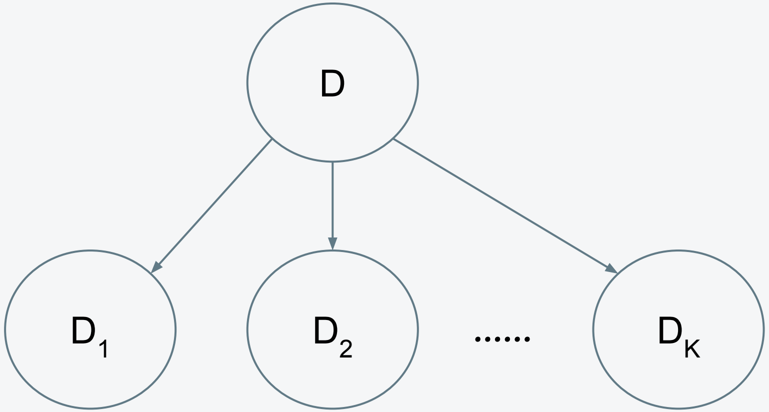


# Bagging

N - number of records (samples, entries)

M - number of features (predictors)

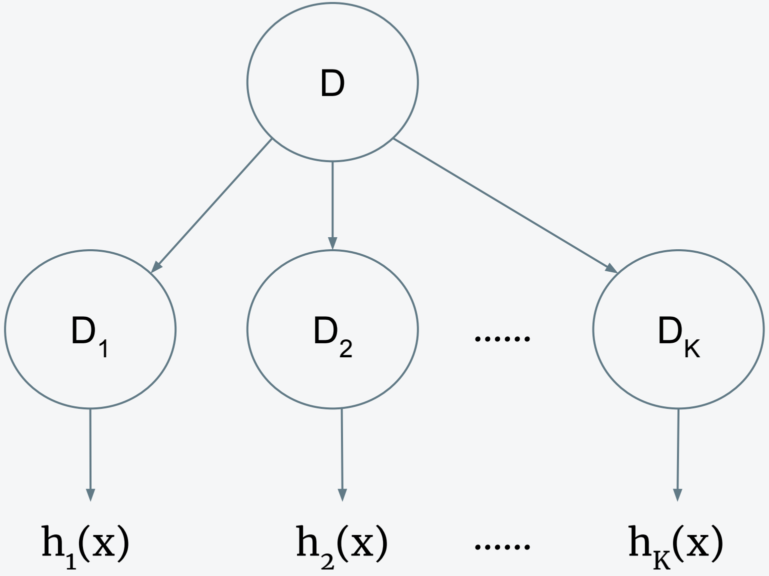
D is N x M




$D_1$... $D_K$ are NxM

(sampled independently with replacement)

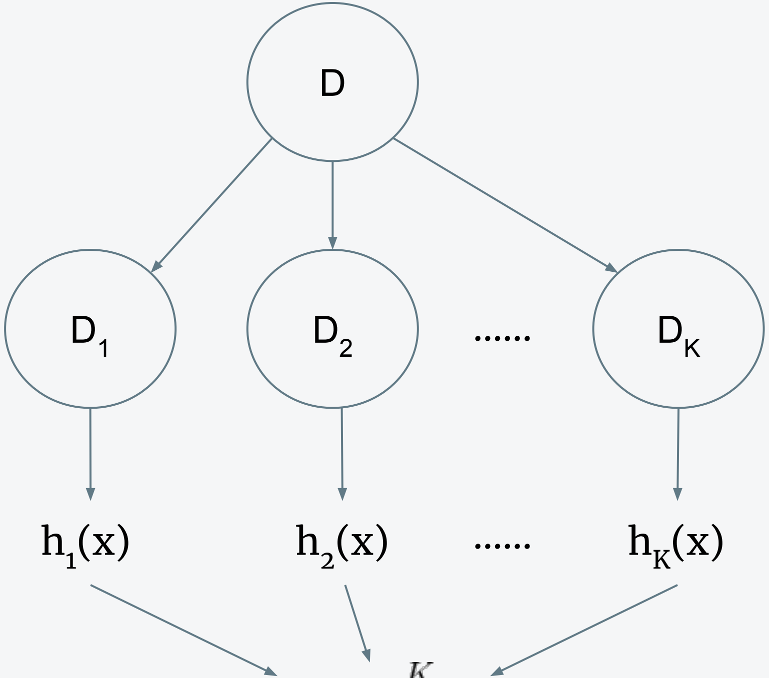
# BAGGING




$D_1 \ldots D_K$ are NxM

(sampled independently with replacement)

# BAGGING




$$h(x) = \frac{1}{K}\sum_{j=1}^{K} h_j(x)$$

$D_1 \ldots D_K$ are NxM

(sampled independently with replacement)

- average of classifiers

# Bagging

Advantages?

Disadvantages?

# Random Forests

Best example of bagging

Algorithm?

Learn a decision tree on each of the "bags" $D_1 \ldots D_K$, except...

# Random Forests

Best example of bagging

Algorithm?

Learn a decision tree on each of the "bags" $D_1 \ldots D_K$, except... at each tree split pick a subset of variables!!!

# Random Forests

Learn a decision tree on each of the samples, except... at each tree-split pick a subset of variables at random (without replacement)!!!

How many new parameters?

# Random Forests

Advantages?

- Fast
- Robust
- Accurate
- Nonlinear
- No need to rescale features
- Few parameters
- Out of bag error
- Feature importance

Disadvantages?

- Treat each dimension independently (e.g. not as good for images)
- May need to prune (extra)

# Out of bag (OOB) error

In principle, don’t need to split data into train and test. Why?

# Feature Importance

Get feature importance for free! How?

## Summary

- There are many classifiers
- Today we learned about decision trees, bagging (not a classifier in itself but helps with overfitting!) and random forests
- Decision tree – simple and usually wrong
- Random forest – very powerful

# Reading Material

Original paper on bagging by Leo Breiman, 1996:

https://link.springer.com/article/10.1007/BF00058655

Original paper on random forests by Leo Breiman, 2001 (quite theoretical):
https://link.springer.com/article/10.1023/A:1010933404324

Machine learning classifiers and fMRI (nice overview of classification itself):

https://www.sciencedirect.com/science/article/pii/S1053811908012263?casa_token=kFvmrR_QnxsAAAAA:H7tk36DYja_88DmfYyIbEEuVGXksXMkZX5LNv3tQoFrKippiZT3UdOfawidRksQoViYqKQlOqKTY