Tutorial 5 – So you’ve decided to use a decision tree...
VC Dimension Review

• How do you calculate VC dimension of **axis-aligned hyperplanes in d dimensions**?
VC Dimension Review

• How do you calculate VC dimension of axis-aligned hyperplanes in d dimensions?
VC Dimension Review

• How do you calculate VC dimension of axis-aligned hyperplanes in $d$ dimensions?
VC Dimension Review

- How do you calculate VC dimension of circles in 2 dimensions
VC Dimension Review

- How do you calculate VC dimension of circles in 2 dimensions
VC Dimension Review

• How do you calculate VC dimension of circles in 2 dimensions
VC Dimension Review

• How about X-classifiers with don’t cares?
VC Dimension Review

• How about X-classifiers with don’t cares?
VC Dimension Review

• How about X-classifiers with don’t cares?
Measuring Hypothesis Spaces

• Rejection cascade of height $h$.
  – feature space is $d$ binary variables
  – target space is $\{\text{reject, accept}\}$
Measuring Hypothesis Spaces

• Rejection cascade of height $h$.
  – feature space is $d$ binary variables
  – target space is \{reject, accept\}
Measuring Hypothesis Spaces

• Rejection cascade of height $h$.
  – feature space is $d$ binary variables
  – target space is \{reject, accept\}

What is the size of this hypothesis space?
DecisionTreeClassifier(self, criterion='gini',
                        splitter='best',
                        max_depth=None,
                        min_samples_split=2,
                        min_samples_leaf=1,
                        max_features=None,
                        random_state=None,
                        min_density=None,
                        compute_importances=None)
DecisionTreeClassifier(self,
criterion='gini',
splitter='best',
max_depth=None,
min_samples_split=2,
min_samples_leaf=1,
max_features=None,
random_state=None,
min_density=None,
compute_importances=None)
Decision trees!

DecisionTreeClassifier(self, criterion='gini', splitter='best', max_depth=None, min_samples_split=2, min_samples_leaf=1, max_features=None, random_state=None, min_density=None, compute_importances=None)
Decision trees!

```
DecisionTreeClassifier(self, criterion='gini',
               splitter='best',
               max_depth=None,
               min_samples_split=2,
               min_samples_leaf=1,
               max_features=None,
               random_state=None,
               min_density=None,
               compute_importances=None)
```
Decision trees!

```
DecisionTreeClassifier(self, criterion='gini',
    splitter='best',
    max_depth=None,
    min_samples_split=2,
    min_samples_leaf=1,
    max_features=None,
    random_state=None,
    min_density=None,
    compute_importances=None)
```
Decision trees!

```
DecisionTreeClassifier(self, criterion='gini',
splitter='best',
max_depth=None,
min_samples_split=2,
min_samples_leaf=1,
max_features=None,
random_state=None,
compute_importances=None)
```
Handling missing values

• Example:
  
  – *Pima Indians Diabetes dataset* (UCI repository)

\[
\begin{bmatrix}
10. & 115. & 0. & 0. & 0. & 0. & 0. & 0.261 & 30. \\
\end{bmatrix}
\]
Handling missing values

• Example:
  – *Pima Indians Diabetes dataset* (UCI repository)
  
  [ 10.  115.  0.  0.  0.  0.  0.  0.261  30. ]

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. **Diastolic blood pressure** (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. **Body mass index** (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
Simple tree...

\[
X[1] \leq 131.5000 \\
gini = 0.457465277778 \\
samples = 576
\]

\[
\text{gini} = 0.3386 \\
samples = 389 \\
value = [ 305.\; 84.]
\]

\[
\text{gini} = 0.4598 \\
samples = 187 \\
value = [ 67.\; 120.]
\]
Pruning

• Hold out a validate set
  – Starting with the leaves, prune any node that can be pruned without increasing the error on the validate set