Applied Machine Learning!!!

W207 Section 9
Rasika Bhalerao
rasikabh@berkeley.edu

Aug 23: Welcome!
Nov 8 and 22: No classes

Schedule

Supervised learning methods

	Sync	Topic
2	Aug 30	Linear Regression / Gradient Descent
3	Sep 6	Feature Engineering Bonus: Naive Bayes
4	Sep 13	Logistic Regression
5	Sep 20	Multiclass classification / Eval Metrics Bonus: Reinforcement learning
6	Sep 27	Neural Networks
7	Oct 4	KNN, Decision Trees, Ensembles

Unsupervised learning methods

	Sync	Topic
8	Oct 11	KMeans and PCA Bonus: LDA
9	Oct 18	Text Embeddings Bonus: Language models
10	Oct 25	CNNs Bonus: GANs
11	Nov 1	EDA, Real data, Baselines
12	Nov 15	Fairness / Ethics
13	Nov 29	Fancy Neural Networks
14	Dec 6	Final Presentations

Assignment Schedule

Due Date	Assignment
Aug 28	HW1
Sep 4	HW2
Sep 11	HW3
Sep 18	HW4
Sep 25	HW5
Oct 2	HW6
Oct 16	Group project baseline
Oct 23	HW8
Nov 6	HW9
Nov 20	HW10
Dec 4	Final project notebook + presentation

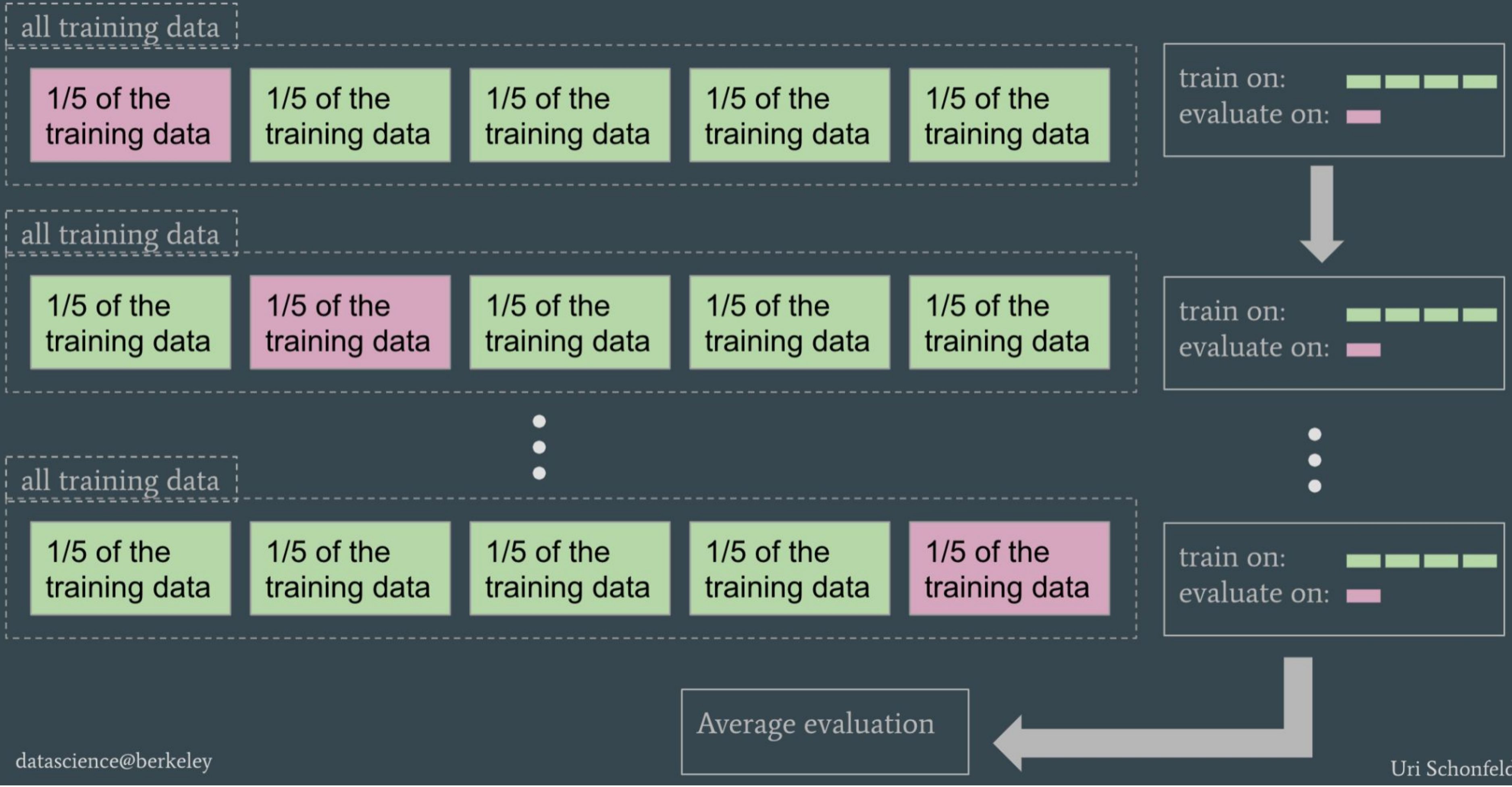
Behavior expectations

- Healthy disagreement is expected
- Be mindful of one another’s schedules
- Be a good listener
- Have fun in a professional manner
- Share related real-world experience
- Ask questions when something is confusing
- Keep it 100 but be respectful
- Be open-minded to new ideas in the real world and when coding
- On time for group meetings

How are final projects going?

Guidelines:
https://docs.google.com/document/d/1R7mIHOtYXKU8vEQzw10uofb_iK3sgimw8iZLWSTzdgg/edit?usp=sharing

Cross validation - 5-fold



K Nearest Neighbors

Quick pseudocode recap of KNN

Use fit:

X = [a, b, c]

Y = [0, 1, 0]

model = KNN(k=5)

model.fit(X, Y)

Implement fit:

```
self.X = X
self.Y = Y
```

Use predict:

X_test = [d, e, f]

Y_predicted = model.predict(X_test)

Implement predict:

```
For x_i in X_test:
    For x_train in self.X:
        dist(x_i, x_train)
    Find the x_trains that yield the k smallest such distances
    Get those x_train's  labels, and the majority is the label for x_i. Call it y_i
Return the list of y_i
```

How would KNN perform if you reused the training data as the test data?

Even or odd k (assuming binary classification)?

View options (at the top) → Annotate

Async Practice Quiz Questions (vote!)

It's possible that a model that perfectly fits all the training data will still generalize well to the test data.	True	False
With length-normalized vectors A and B, Euclidean distance e(A,B) is related to dot product distance d(A,B) by: e(A,B)^2 / 2 = 1-d(A,B).	True	False

What other distance metrics are there?

- L1, L2, ...
- Euclidean distance
- Manhattan distance
- Hamming distance
- Levenshtein distance
- Jaccard index
- Cosine similarity
- Wordnet
- Create your own!

Decision Trees

Review: How do we use a trained decision tree to predict?

Input x

Start with the first node

While the node is not a leaf

    Follow the path on the node

Return the prediction associated with that leaf

Review: How do we train a decision tree?

Function (x,y pairs):

    If the dataset is empty:

        Assign Y = majority of the parent node’s Y

    If all y are the same:

        leaf node

    Split on feature that gives highest information gain

    For each child node created:

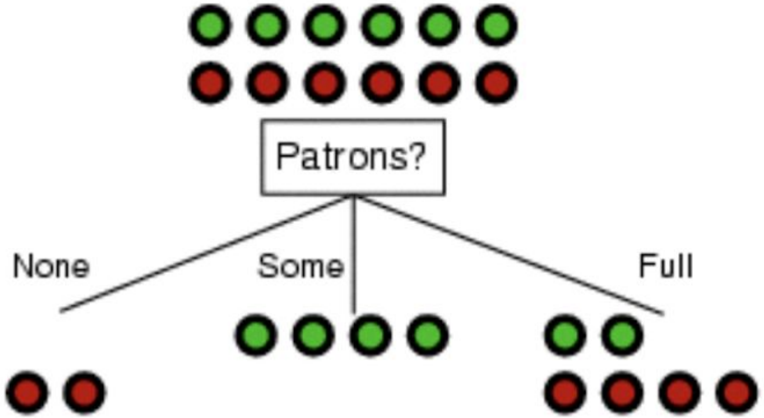
        Call this function with the subset of the data in that child node

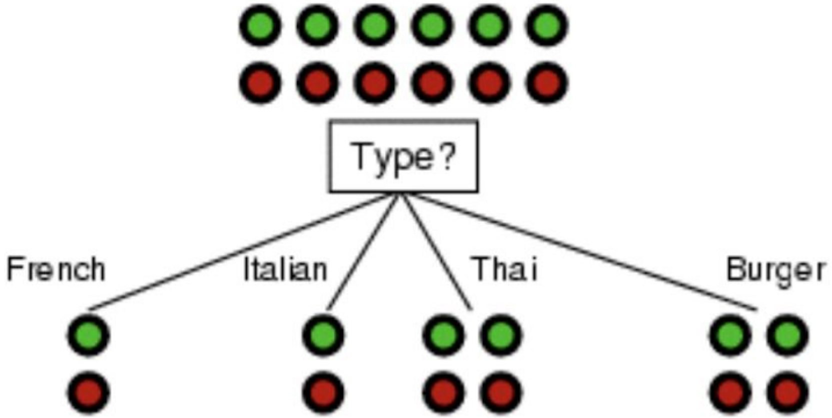
Choosing an attribute

Example: deciding whether or not we will wait for a table at a restaurant

Green: we have waited for it before
Red: we have not waited for it before

Which feature gives more information?

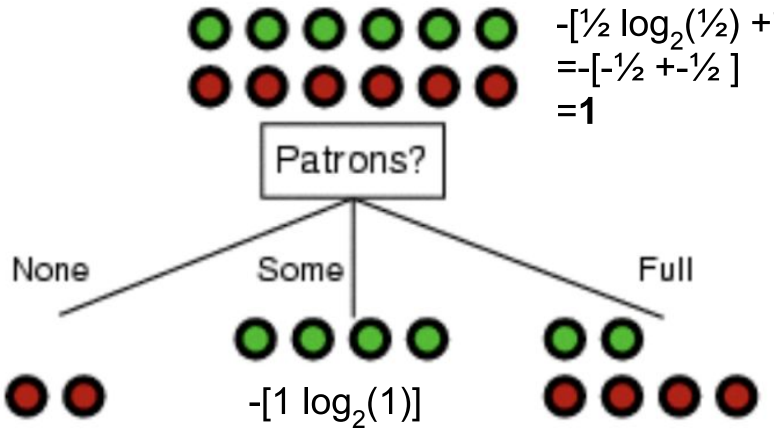




Choosing an attribute: entropy

Entropy: measure of “uncertainty” in data

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

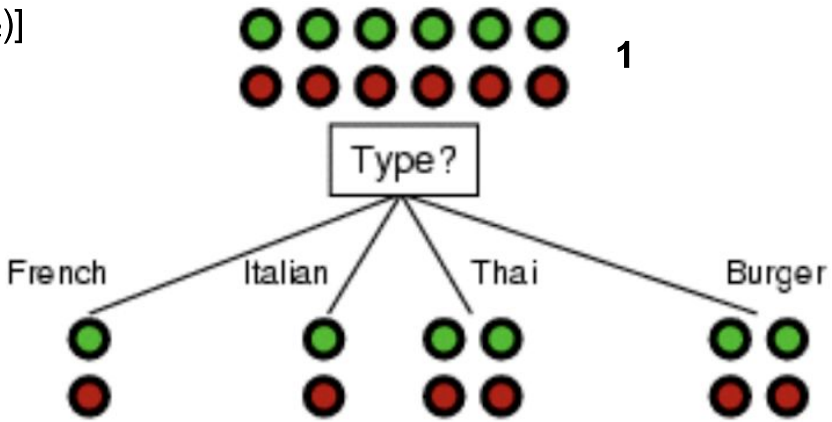


$-[\frac{1}{2} \log_2(\frac{1}{2}) + \frac{1}{2} \log_2(\frac{1}{2})]$
$= -[-\frac{1}{2} + -\frac{1}{2}]$
=1

$-[1 \log_2(1)]$
=0

$-[1 \log_2(1)]$
=0

$-[\frac{1}{3} \log_2(\frac{1}{3}) + \frac{2}{3} \log_2(\frac{2}{3})]$
$= -[-0.53 + -0.39]$
=0.92



1

$-[\frac{1}{2} \log_2(\frac{1}{2}) + \frac{1}{2} \log_2(\frac{1}{2})]$
$= -[-\frac{1}{2} + -\frac{1}{2}]$
=1
For all four of these leaves

Attribute selection measure: Information gain

- Choose the attribute with **highest information gain** (reduces the most entropy)
- Let p_i be the probability that an arbitrary element of D belongs to class C_i, estimated by $|C_i \cap D|/|D|$
- Information (entropy) in D: $Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$
- Information by splitting D on attribute A into v partitions:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

- Information gained after using A to split D: $Gain(A) = Info(D) - Info_A(D)$

Attribute selection: information gain

Class P: buys_computer = "yes"
Class N: buys_computer = "no"

$$Info(D) = I(9,5) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940$$

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
31…40	4	0	0
>40	3	2	0.971

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

$$Info_{age}(D) = \frac{5}{14}I(2,3) + \frac{4}{14}I(4,0) + \frac{5}{14}I(3,2) = 0.694$$

$\frac{5}{14}I(2,3)$ means "age <=30" has 5 out of 14 samples, with 2 yes'es and 3 no's. Hence

$$Gain(age) = Info(D) - Info_{age}(D) = 0.246$$

Similarly,

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

$$Gain(credit\_rating) = 0.048$$

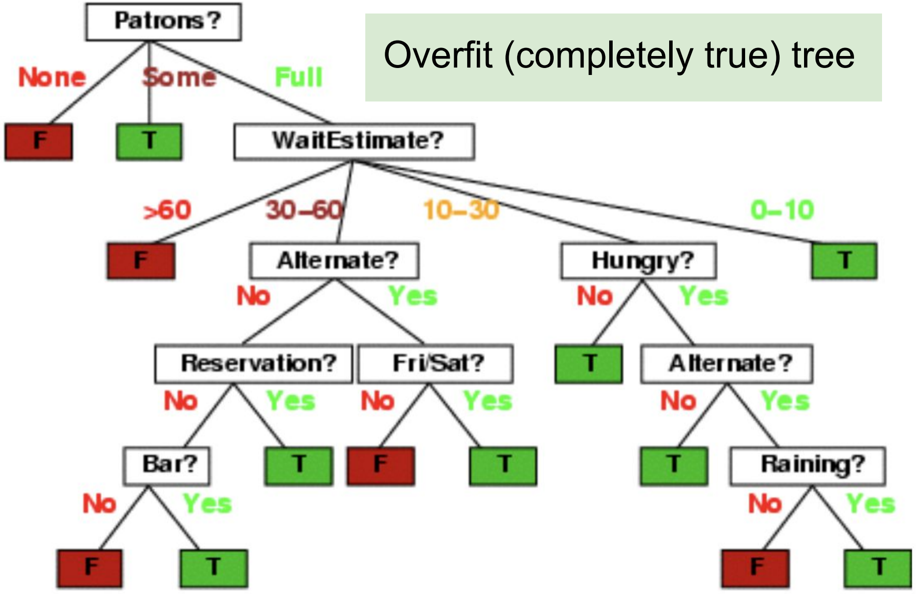
Should we keep a feature that we already used for a split to use in its subtree?

Overfitting decision trees

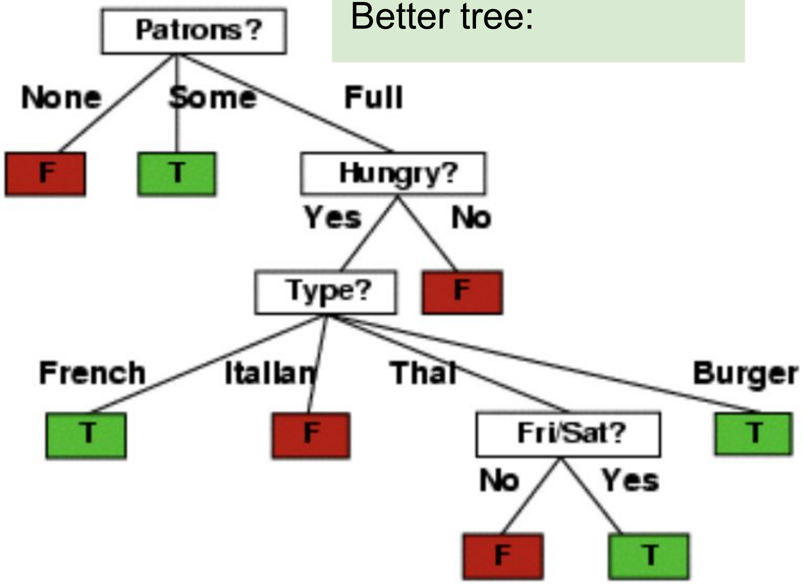
- What is wrong here?
- How can we fix it?

Example	Attributes										Target
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*Wait*
X_1	T	F	F	T	Some	$$$	F	T	French	0–10	T
X_2	T	F	F	T	Full	$	F	F	Thai	30–60	F
X_3	F	T	F	F	Some	$	F	F	Burger	0–10	T
X_4	T	F	T	T	Full	$	F	F	Thai	10–30	T
X_5	T	F	T	F	Full	$$$	F	T	French	>60	F
X_6	F	T	F	T	Some	$$	T	T	Italian	0–10	T
X_7	F	T	F	F	None	$	T	F	Burger	0–10	F
X_8	F	F	F	T	Some	$$	T	T	Thai	0–10	T
X_9	F	T	T	F	Full	$	T	F	Burger	>60	F
X_{10}	T	T	T	T	Full	$$$	F	T	Italian	10–30	F
X_{11}	F	F	F	F	None	$	F	F	Thai	0–10	F
X_{12}	T	T	T	T	Full	$	F	F	Burger	30–60	T

Overfit (completely true) tree



Better tree:



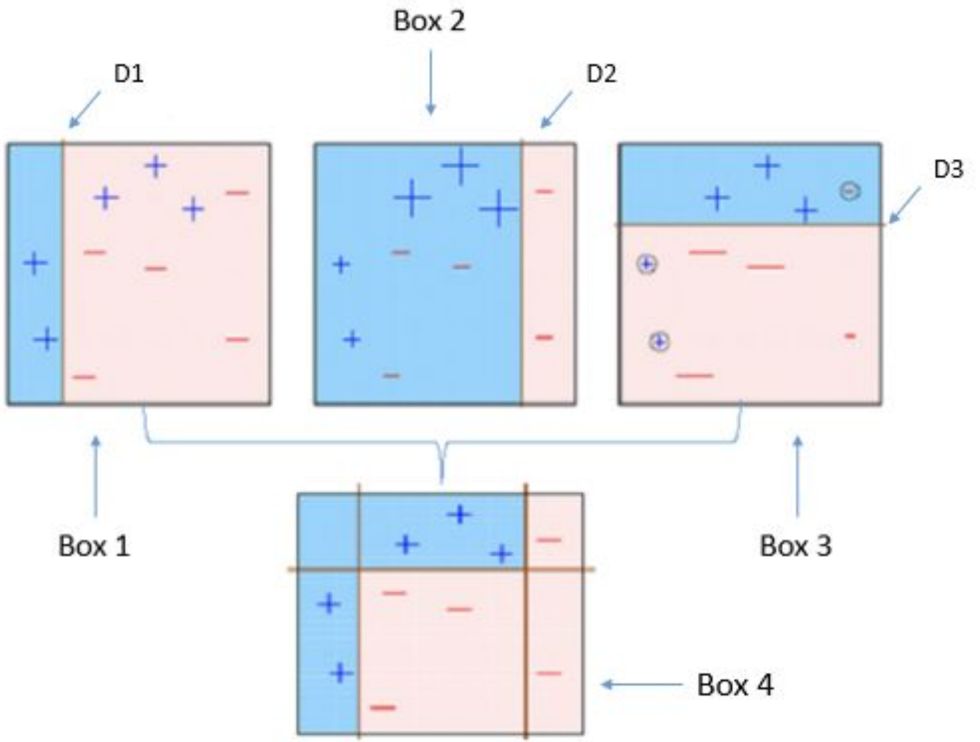
View options (at the top) → Annotate

Async Practice Quiz Questions (vote!)

A decision tree cannot learn interactions between features.	True	False
The decision tree algorithm finds the optimal tree (the smallest tree that explains the training data).	True	False
A decision tree will always be a balanced binary tree.	True	False
Bagging and boosting can be used with any classifier.	True	False

What is Boosting?

- (How is it different from Bagging?)



Notebook!

To access later:
https://github.com/MIDS-W207/rasikabh/blob/main/live_sessions/Week7.ipynb

Also, if you want last semester's assignments:
https://github.com/MIDS-W207/coursework