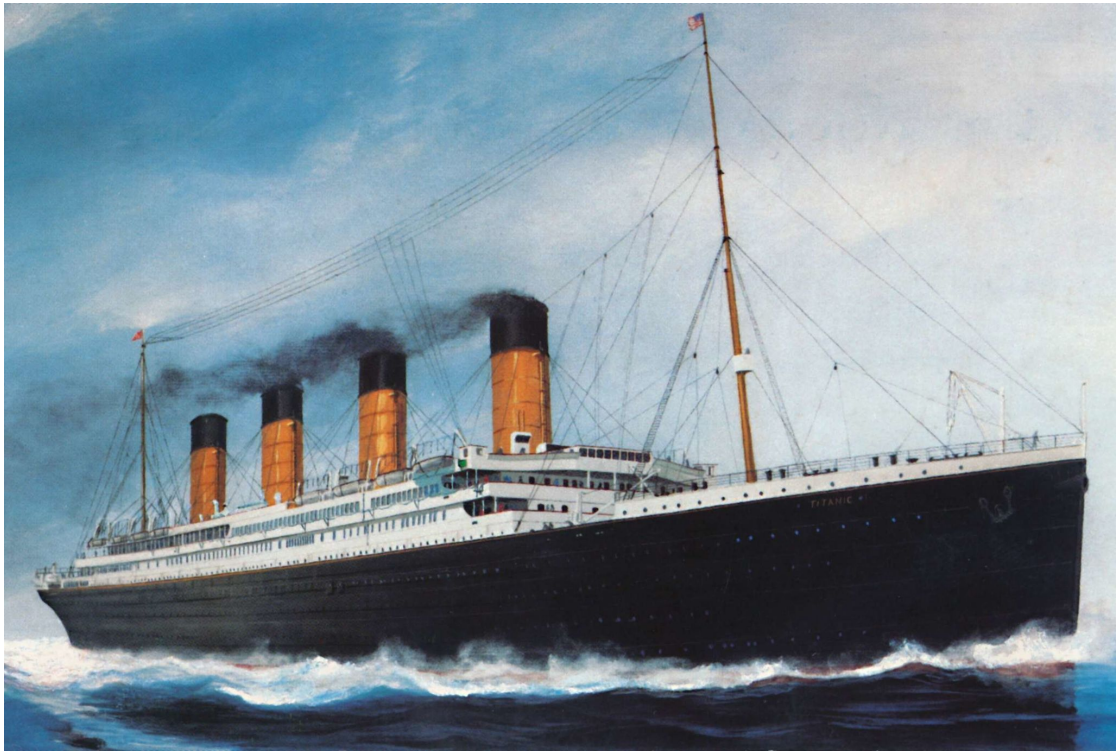
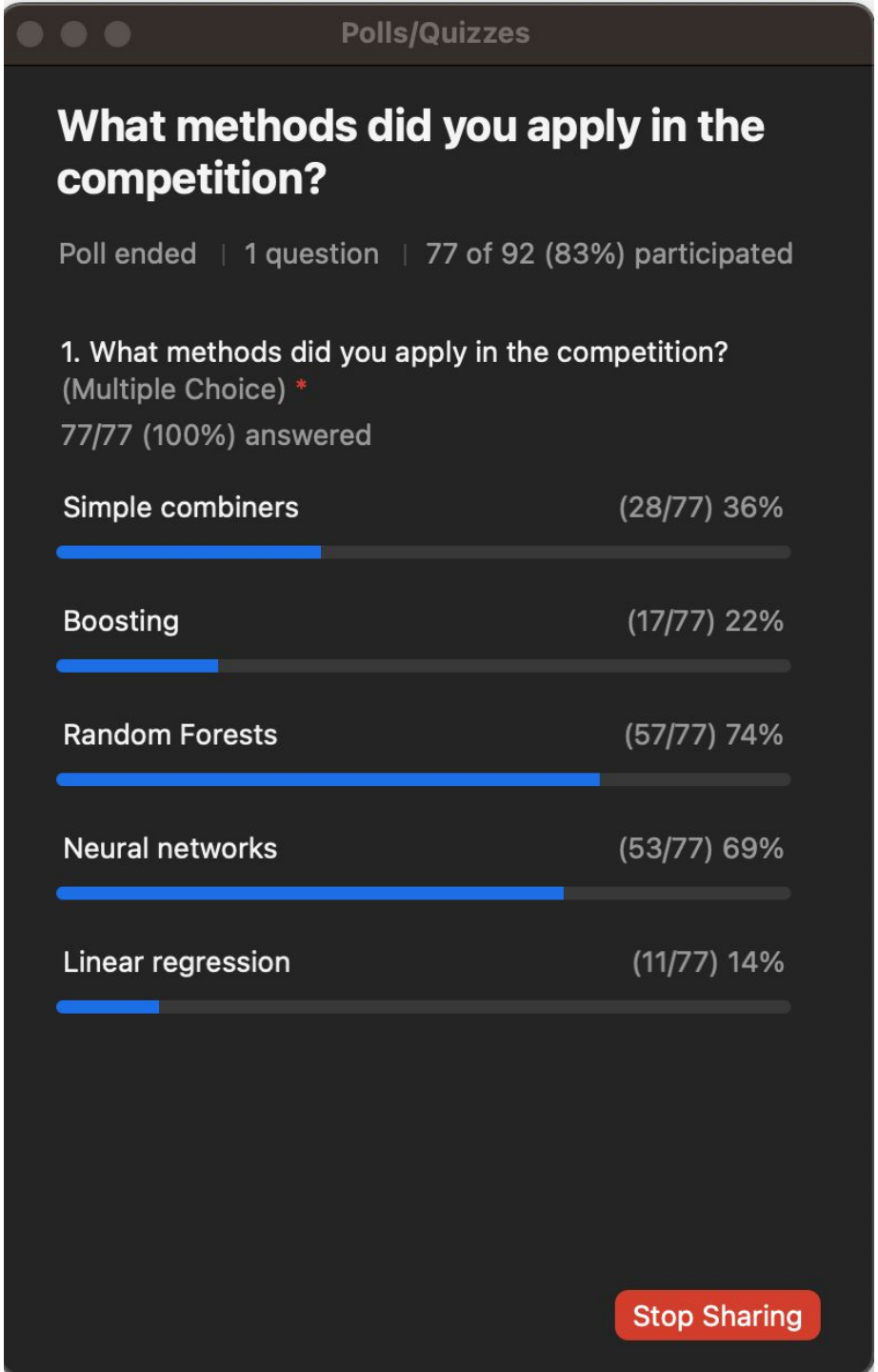


Kaggle Competition Lab Reflections


Polls/Quizzes
What methods did you apply in the competition?
Poll ended | 1 question | 77 of 92 (83%) participated
1. What methods did you apply in the competition? (Multiple Choice) *
77/77 (100%) answered
Simple combiners (28/77) 36%
Boosting (17/77) 22%
Random Forests (57/77) 74%
Neural networks (53/77) 69%
Linear regression (11/77) 14%
Stop Sharing

Midterm Review

Data Mining & Analytics

Midterm

- Wednesday, October **19th** in-class
- The Midterm will be on bCourses (like a quiz)
- The link to the Midterm will be posted before class
  - 1. All exams must be turned in no later than 5:00pm
  - 2. You will have no more than 2 hours to complete the exam
  - 3. To receive the full 2 hours, you must start between 2pm and 3pm
  - 4. If starting at 2pm, for example, your exam will be due at 4pm
  - 5. If starting at 3:30pm, for example, your exam will be due at 5pm
  - 6. There are 19 questions, worth 53 points total (8 points extra credit + 15%)
  - 7. The exam is open book/note
  - 8. No communication is allowed during the test

Topics

- Data transformation (pre-processing)
- Clustering (k-means)
- Classification (decision trees, neural nets)
- Model evaluation (cross-validation, error metrics)
- Combining classifiers (ensembles)

Data transformation (pre-processing)

Subtopics

- Feature engineering (pandas)
- Representing data to fit the prediction task
- Normalization

  e.g., Z-score:

Formula to find population mean

$$\mu = \frac{\sum x}{n}$$

Formula to find population standard deviation

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{n}}$$

Formula to find the **z-score**

$$z\ score = \frac{(x - \mu)}{\sigma}$$

Example Exam Question

1. The below table shows a log file, questions.csv, of students answering quiz questions. Each row in this table represents a single student's answer to a single quiz question. (4pts total)

Quiz Question ID	Student Name	Answer Score
DATA_TRANSFORM1	Geoff Hinton	1
KMEANS2	Geoff Hinton	0
DATA_TRANSFORM1	Grace Hopper	0
KMEANS2	Grace Hopper	1
DATA_TRANSFORM1	Dennis Hopper	0
KMEANS2	Dennis Hopper	0
SPECTRAL3	Dennis Hopper	1

Table 1. Original question answer table

(a) Translate the above dataset from a one row per question answer dataset to a one row per student dataset. Fill in all of the blank cells of the table below. (2pts)

Student Name	Student % Correct	Z-score of % Correct	# of questions answered

Table 2. Transformed student table

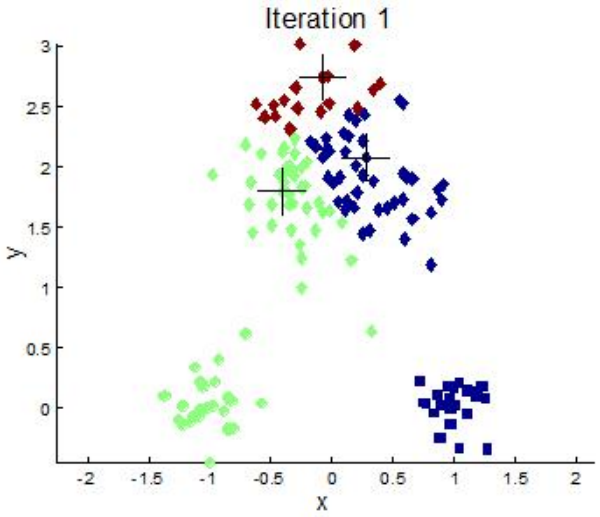
(b) Given a pandas dataframe, **df**, of the original questions.csv file, write a statement that would output percent correct for Geoff Hinton. (2pt)

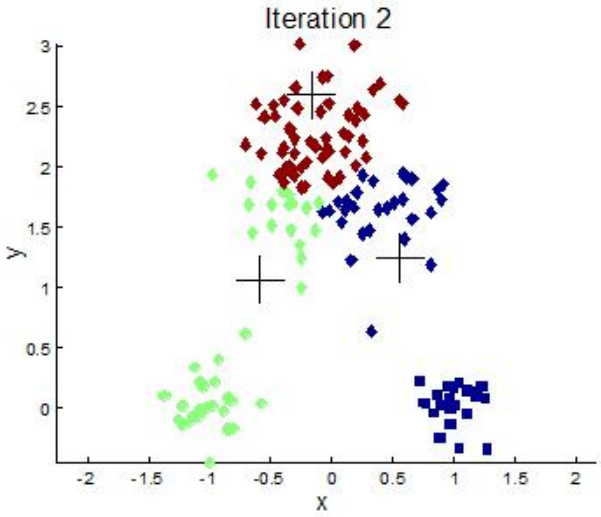
Clustering (k-means)

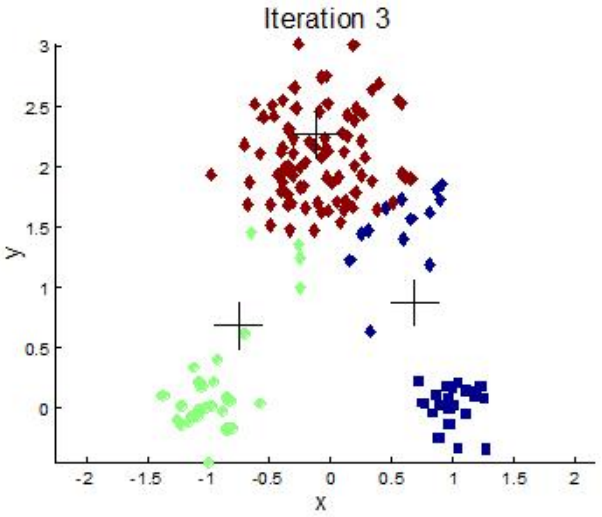
Subtopics

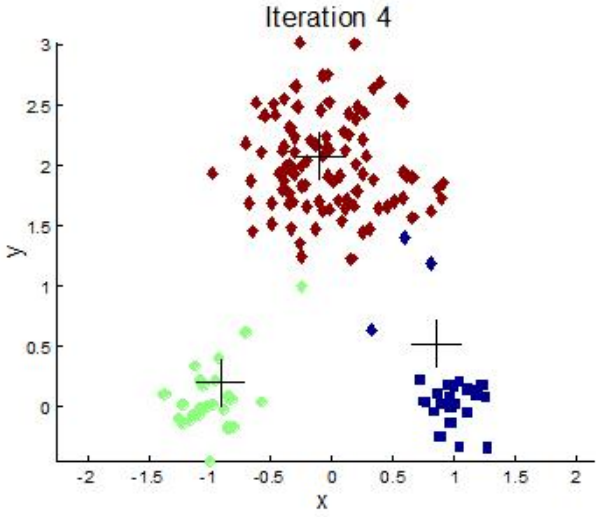
- Types of clustering methods
- Measures of cluster goodness (SSE, silhouette score)
- The k-means algorithm
- Ways of choosing K (elbow method)

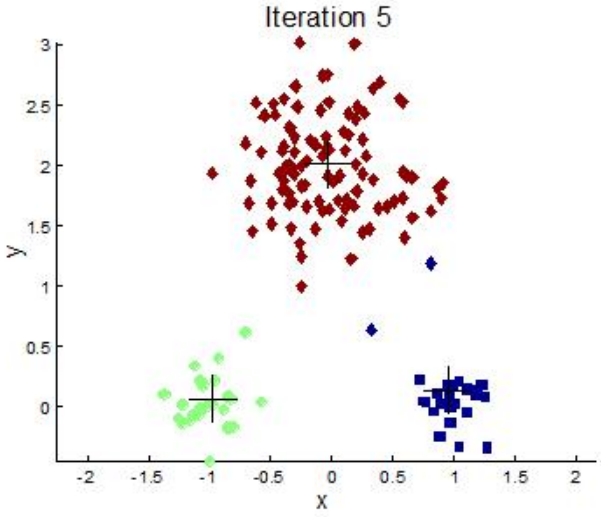
Clustering (k-means)

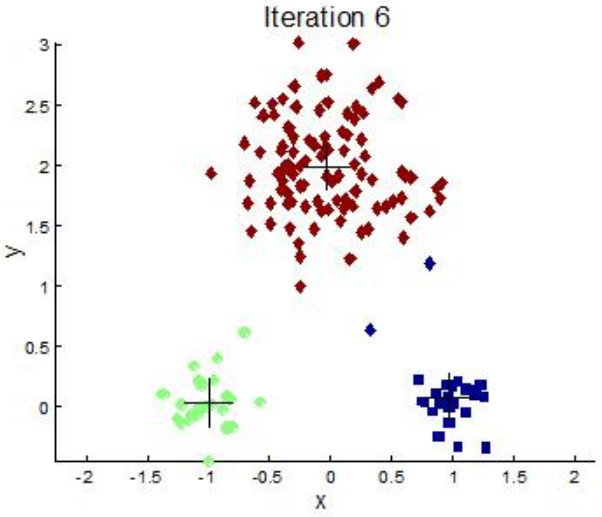




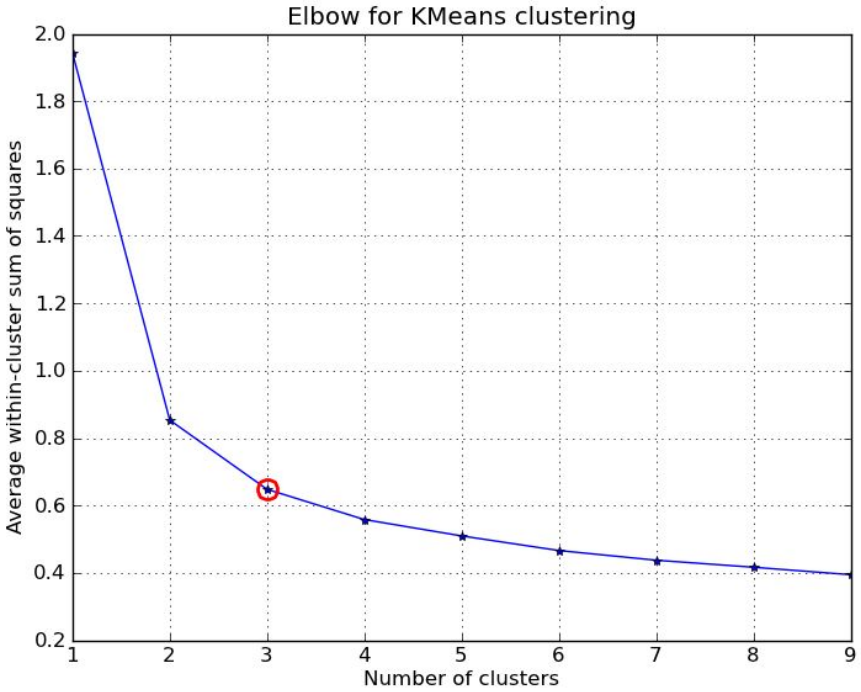








Clustering: Elbow Method



$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(\boldsymbol{p}, \boldsymbol{c_i})^2,$$

Sum of Squared Errors

Clustering: SSE, Silhouette Score

Within-cluster Variance / Sum of Squared Errors

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(\boldsymbol{p}, \boldsymbol{c_i})^2,$$

Silhouette Score (take the avg. of all s(o)) - every data point

For each data point o in Ci:

$$\boldsymbol{a(o)} = \frac{\sum_{\boldsymbol{o'} \in C_i, \boldsymbol{o} \neq \boldsymbol{o'}} dist(\boldsymbol{o}, \boldsymbol{o'})}{|C_i| - 1}$$

$$\boldsymbol{b(o)} = \min_{C_j : 1 \leq j \leq k, j \neq i} \left\{ \frac{\sum_{\boldsymbol{o'} \in C_j} dist(\boldsymbol{o}, \boldsymbol{o'})}{|C_j|} \right\}$$

$$\boldsymbol{s(o)} = \frac{b(\boldsymbol{o}) - a(\boldsymbol{o})}{\max\{a(\boldsymbol{o}), b(\boldsymbol{o})\}}.$$

Example Exam Question

Given a clustering assignment on four 2-dimensional points:

Pt	Feature 1	Feature 2	Cluster
A	2	2	1
B	0	2	1
C	-4	-1	2
D	-3	-2	2

(a) Calculate the silhouette coefficient for point B.

(b) If this assignment is obtained right after an iteration of K-means clustering (which may or may not have terminated), do you think the assignment will change in later iterations? Why or why not?

Classification (decision trees)

Subtopics

- Characterizing purity (Gini/Info)
- Splitting based on features to improve purity (trees)
- Improving the generalizability of training trees (pruning)

Example Exam Question

Pt	Feature	Label
A	7	0
B	10	1
C	4	0
D	10	0
E	16	1
F	9	1

(a) Calculate the gini index of the Dataset

(b) If we split on 8, what is the overall gini index after splitting?

(c) If we split on 13, what is the overall gini index after splitting?

Classification (neural networks)

Subtopics

- Feed forward neural networks

  Input layer, hidden layer, output layer, weights, bias
- Backpropagation (conceptual)
- Activation functions

  Logistic, relu, softmax
- Additional details

  Epoch, batch size, stopping criteria

Example Exam Question

9. The number of nodes in the input layer of a neural network is equal to which of the following? (circle one) *(1 pt)*

- the number of rows in the dataset
- the number of features (columns)
- the number of bytes used to represent the feature space

10. Nodes in which layers of a multi-layer perceptron network contain activation functions? *(1 pt)*

Evaluation of models

Subtopics

- Metrics (confusion matrix based & continuous)
- Training, validation, and testing sets
- Cross-validation
- Model selection

Error Metrics

Measure	*Formula*
accuracy, recognition rate	$\frac{TP+TN}{P+N}$
error rate, misclassification rate	$\frac{FP+FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$
F, F_1, F-score, harmonic mean of precision and recall	$\frac{2 \times precision \times recall}{precision + recall}$

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Examples from lecture

predicted	actual
0	0
0	1
1	0
1	1
1	1

predicted	actual
0.25	0
0.45	1
0.66	0
0.71	1

Example Exam Question

For each of the tasks below, explain which algorithm(s) and error metric(s) you would use and why.

(a) Predict GPA given students' department, credits taken, study hours, etc.

(b) Predict if a Twitter user is liberal or conservative.

(c) Predict lung cancer from chest X-rays.

Combining classifiers (ensembles)

Subtopics

- Simple combiners
- Bagging (e.g., random forests)
- Boosting (Adaboost)
- Blending/Stacking

Example Exam Question

Determine True or False for each of the following statements:

(a) Ensemble methods are never the cause of overfit.

(b) Hyperparameters of random forests include (but not limited to) number of trees, percentage of rows sampled, and max depth.

(c) Blending uses cross-validation while stacking uses a holdout validation.

Break-out groups

(self-select after break)

- Clustering & Preprocessing
- Prediction/Classification Models
- Cross-validation & Metrics
- Ensembling

Prof. Pardos will stay in the main room for general Q & A