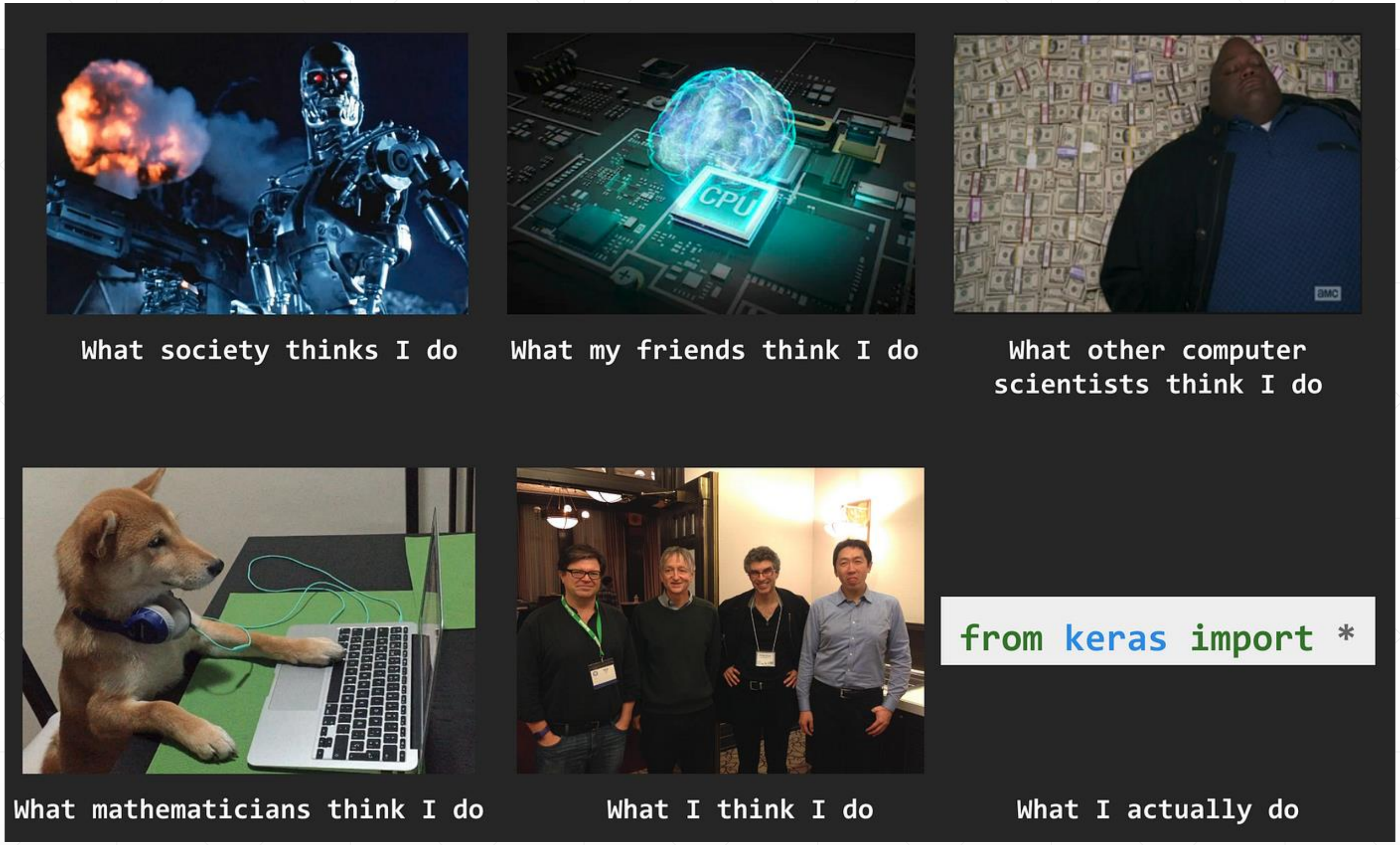


Source

INFO251 – Applied Machine Learning

Lab 14

Satej Soman, Suraj R. Nair **(based on previous material by Emily Aiken & Josh Blumenstock)**

Announcements

- Please fill out the course evaluation (current response rate is 10% ☹)
  - https://course-evaluations.berkeley.edu/
- **PS7** due May 5
- **Quiz 2** on Thursday May 1

Let us know via email or Edstem or in person if you have a DSP accommodation / time conflict RIGHT AFTER LAB

Agenda

- PCA visualized
- Topics covered in AML
- ML algorithms review
- Practice quiz questions

Topics covered in AML

1. Causal inference
- Linear regression
- Fixed effects and panel data
- Instrumental variables
- Regression discontinuity

2. Supervised Learning, Part 1
- K-nearest neighbors
- Linear regression
- Logistic regression
- Ridge and LASSO
- Support vector machines

3. Loss Functions & Optimization
- Mean squared error
- Logistic loss
- Cross entropy loss
- Loss functions w. regularization
- Gradient Descent

4. Supervised Learning, Part 2
- Naïve Bayes
- Decision Trees
- Random Forests
- Gradient Boosting

5. Neural Networks and LLMs
- Perceptron
- Fully Connected Networks
- Autoencoders
- Convolutional Neural Networks
- Recurrent Neural Networks / LSTM
- Embeddings
- Attention, self-attention & multi-head attention
- Transformers (LLMs/ vision)
- Pre-trained models
- Fine-tuning

7. Bias & Fairness in ML
- ML Failures
- Detecting bias
- Ameliorating bias
- p% - rule

8. Unsupervised Learning
- K-means clustering
- Hierarchical clustering
- Dimensionality reduction
- Principal components analysis

9. Practical ML
- Train-test splits
- Regularization
- Cross validation
- Feature engineering
- Missing data
- Feature scaling
- Imbalanced data
- Overfitting
- Bias-variance trade-off
- Interpretability
- Error + ablative analysis

Python programming tools covered in AML

Tool	Purpose
`numpy`	Coding up algorithms, vectorized computation
`pandas`	Storing real-world tabular data
`matplotlib, seaborn`	Visualization
`statsmodels`	Linear regression for causal inference
`scikit-learn`	Supervised and unsupervised learning pipelines
`xgboost`	Gradient boosting models
`pytorch, transformers`	Neural networks
`imbalanced-learn`	Handling imbalanced data

ML Algorithms Summary: Linear Models

Algorithm	Applications	Hyperparameters	Description	Pros	Cons
Linear Regression	Regression	--	Prediction for observation is linear combination of features, weights determined via optimization (gradient descent).		
LASSO/Ridge Regression	Regression	• Regularization (L1 or L2) • Regularization strength (lambda)	Regularized linear regression, penalizing size of weight vector		
Logistic Regression	Classification	• Regularization (L1 or L2) • Regularization strength (lambda)	Regression optimizing logistic loss to produce calibrated class probabilities		
Support Vector Machines	Classification	• Regularization strength (C)	Maximize margin around separating hyperplane, with penalties for misclassification		

ML Algorithms Summary: Linear Models

Algorithm	Applications	Hyperparameters	Description	Pros	Cons
Linear Regression	Regression	--	Prediction for observation is linear combination of features, weights determined via optimization (gradient descent).	• Directly interpretable coefficients • Closed form solution • Scalable	• Overly simplistic model • Cannot learn nonlinear decision boundaries • Overfitting
LASSO/Ridge Regression	Regression	• Regularization (L1 or L2) • Regularization strength (lambda)	Regularized linear regression, penalizing size of weight vector	• Reduces overfitting • Optimal regularization determined through cross validation • Feature selection (Lasso only)	• Cannot learn nonlinear decision boundaries
Logistic Regression	Classification	• Regularization (L1 or L2) • Regularization strength (lambda)	Regression optimizing logistic loss to produce calibrated class probabilities	• Directly interpretable coefficients • Scalable • Option to add regularization	• Cannot learn nonlinear decision boundaries
Support Vector Machines	Classification	• Regularization strength (C)	Maximize margin around separating hyperplane, with penalties for misclassification	• Easy to regularize • Works with kernels	• Performs badly when data not linearly separable • Linear decision boundary only • No class probabilities

ML Algorithms Summary: Nonlinear Models

Algorithm	Applications	Hyperparameters	Description	Pros	Cons
K-Nearest Neighbors	Regression, Classification	• Neighbors (K) • Distance metric	Prediction for observation is average of target value for K closest observations in training set.		
Naïve Bayes	Classification, text data	• Additive smoothing parameter	MAP estimate for most likely class given the data (features)		
Decision Trees	Regression, Classification	• Maximum depth • Minimum samples in leaves	Recursively grow a tree splitting on a feature value at each node		
Random Forests	Regression, Classification	• Maximum depth • Minimum samples in leaves • Number of trees	Ensemble method aggregating multiple trees via averaging (regression) or voting (classification)		
Gradient Boosting	Regression, Classification	• All of above • Learning rate	Ensemble method where trees built sequentially based on where previous trees performed badly		

ML Algorithms Summary: Nonlinear Models

Algorithm	Applications	Hyperparameters	Description	Pros	Cons
K-Nearest Neighbors	Regression, Classification	• Neighbors (K) • Distance metric	Prediction for observation is average of target value for K closest observations in training set.	• Simple, intuitive, interpretable • No training required	• Slow • Must choose a good distance metric
Naïve Bayes	Classification, text data	• Additive smoothing parameter	MAP estimate for most likely class given the data (features)	• Generative model • Easy, parallelizable estimation	• Conditional independence assumption violated
Decision Trees	Regression, Classification	• Maximum depth • Minimum samples in leaves	Recursively grow a tree splitting on a feature value at each node	• Can learn nonlinear decision boundaries • Most interpretable model	• Prone to overfitting
Random Forests	Regression, Classification	• Maximum depth • Minimum samples in leaves • Number of trees	Ensemble method aggregating multiple trees via averaging (regression) or voting (classification)	• Can learn highly nonlinear decision boundaries • Can cross validate a number of parameters • Parallelizable	• Difficult to interpret
Gradient Boosting	Regression, Classification	• All of above • Learning rate	Ensemble method where trees built sequentially based on where previous trees performed badly	• Can learn highly nonlinear decision boundaries • Typically more accurate than random forests	• Difficult to interpret • Less parallelizable

ML Algorithms Summary: Neural Networks

Algorithm	Applications	Hyperparameters (Architecture)	Description	Pros	Cons
Fully Connected Neural Network	Tabular data	• Number of hidden layers • Number of nodes in hidden layers • Activation functions • Regularization/dropout	All nodes in layer of network connected to all nodes in next layer.		
Convolutional Neural Network	Image data, graph data	• Filter size and stride • Pooling • Conv + pool stacks / blocks • Number of fully connected layers at the end	Convolutional layers learn spatial dependencies, pooling layers reduce image size/complexity.		
Recurrent Neural Network	Time series data, text data	• Network structure (RNN, LSTM, GRU) • Regularization	Recurrent connections allow information to be passed from one input to the next		
Transformers	Time series, text data	• Embedding dimensions • KVQ dimensions • Number of attention heads	Modeling semantic embeddings and positions of tokens allows for effective sequence-to-sequence modeling		

Common hyperparameters when training / fine-tuning NN: batch size, number of epochs, optimizer and learning rate

ML Algorithms Summary: Neural Networks

Algorithm	Applications	Hyperparameters (Architecture)	Description	Pros	Cons
Fully Connected Neural Network	Tabular data	• Number of hidden layers • Number of nodes in hidden layers • Activation functions • Regularization/dropout	All nodes in layer of network connected to all nodes in next layer.	• Faster to train (than more complex network) • Work well for tabular data	• Expensive to train • Must choose a good distance metric • Overfitting
Convolutional Neural Network	Image data, graph data	• Filter size and stride • Pooling • Conv + pool stacks / blocks • Number of fully connected layers at the end	Convolutional layers learn spatial dependencies, pooling layers reduce image size/complexity.	• Very good at learning dependencies in spatial data	• Expensive to train • Overfitting
Recurrent Neural Network	Time series data, text data	• Network structure (RNN, LSTM, GRU) • Regularization	Recurrent connections allow information to be passed from one input to the next	• Very good at learning temporal dependencies	• Long-term dependencies lost in standard RNNs
Transformers	Time series, text data	• Embedding dimensions • KVQ dimensions • Number of attention heads	Modeling semantic embeddings and positions of tokens allows for effective sequence-to-sequence modeling	• Parallelized inference without embedding bottleneck • Long-range dependencies captures	• Data- and compute-intensive

Common hyperparameters when training / fine-tuning NN: batch size, number of epochs, optimizer and learning rate

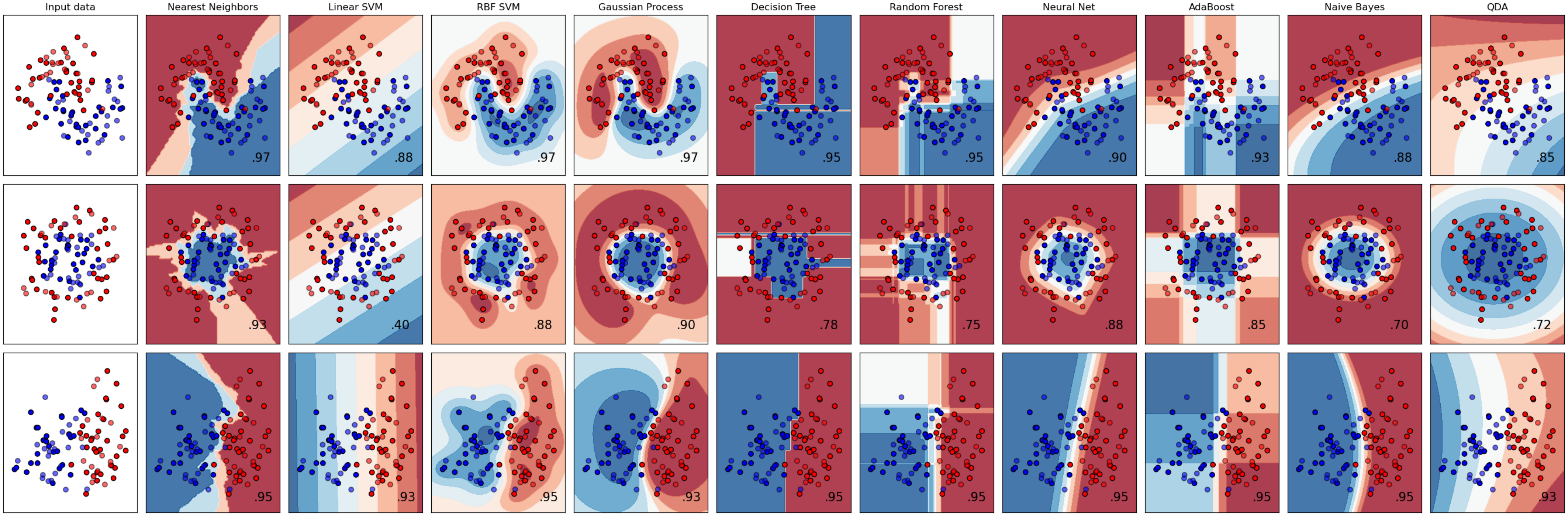
ML Algorithms Summary: Unsupervised Methods

Algorithm	Applications	Hyperparameters	Description	Pros	Cons
K-means clustering	Unsupervised Learning (Clustering)	• Distance metric • Number of clusters	Assign cluster centers randomly. Then, repeat until converged: assign all observations to closest cluster center, assign cluster centers as mean of observations in cluster.		
Hierarchical Clustering	Unsupervised Learning (Clustering)	• Distance metric • Linkage function	Agglomerative clustering starts with all observations in single clusters and links nearby clusters recursively, divisive clustering starts with all observations in single cluster and splits clusters recursively.		
Principal Components Analysis	Unsupervised Learning (Dimensionality Reduction)	• Number of components	Project data into lower dimensional subspace defined by principal components, where components maximize variation explained from original data and are all orthogonal.		

ML Algorithms Summary: Unsupervised Methods

Algorithm	Applications	Hyperparameters	Description	Pros	Cons
K-means clustering	Unsupervised Learning (Clustering)	• Distance metric • Number of clusters	Assign cluster centers randomly. Then, repeat until converged: assign all observations to closest cluster center, assign cluster centers as mean of observations in cluster.	• Guaranteed to converge • Intuitive	• Spherical clusters • All observations assigned to single cluster • Not always clear how to pick number of clusters • Sensitive to random initialization
Hierarchical Clustering	Unsupervised Learning (Clustering)	• Distance metric • Linkage function	Agglomerative clustering starts with all observations in single clusters and links nearby clusters recursively, divisive clustering starts with all observations in single cluster and splits clusters recursively.	• Doesn't require number of clusters (k)	• Expensive to compute • Sensitive to linkage function • Sensitive to random initialization
Principal Components Analysis	Unsupervised Learning (Dimensionality Reduction)	• Number of components	Project data into lower dimensional subspace defined by principal components, where components maximize variation explained from original data and are all orthogonal.	• Very computationally efficient • Can reduce overfitting for supervised learning	• Information may be lost in lower dimensional embedding (check variance explained)

ML Algorithms Summary: Decision Boundaries



scikit-learn.org

Practical ML – Incomplete list of resources

- Best Practices for ML Engineering (Google guide)
- Andrew Ng's slides on applying ML
- Writing ML code with sklearn
  - Hyperparameter tuning -- tips
  - Common pitfalls and recommended practices
  - Optimizing computational performance
- Tuning gradient descent (scroll to the section on tricks of the trade)
- Deep learning
  - Common CNN architectures
  - Analysis of Deep Learning Models for Practical Applications
  - CS231n notes on CNN practicalities and computational considerations ("In practice: use whatever works best on ImageNet" :P)
  - Deep learning Tuning Playbook (Google Guide)
- A practical guide to LLMs
- Transformers - Tips and tricks for training
- Good Data Analysis (Google guide)

Practice Quiz Question 1

Linear regression

Using the California Housing Dataset, you run a linear regression to predict the median house value of a neighborhood based on whether it is adjacent to the ocean (Ocean) and the age of the house (Age). The results are at right. Which of the following are true?

	Coefficient (in 1000s)	95% confidence interval
Intercept	500	[455, 545]
Ocean	250	[225, 275]
Age	-10.3	[-30.7, 10.1]

(A) A new house (age = 0) which is far from the Ocean would have an expected median housing value of $500,000
(B) For a 10 year increase in age, housing value drops by ~$100,000
(C) Being next to the Ocean decreases housing value
(D) Both Ocean and Age are statistically significant predictors at a 0.05 level

Practice Quiz Question 1

Linear regression

Using the California Housing Dataset, you run a linear regression to predict the median house value of a neighborhood based on whether it is adjacent to the ocean (Ocean) and the age of the house (Age). The results are at right. Which of the following are true?

	Coefficient (in 1000s)	95% confidence interval
Intercept	500	[455, 545]
Ocean	250	[225, 275]
Age	-10.3	[-30.7, 10.1]

(A) A new house (age = 0) which is far from the Ocean would have an expected median housing value of $500,000
(B) For a 10 year increase in age, housing value drops by ~$100,000
(C) Being next to the Ocean decreases housing value
(D) Both Ocean and Age are statistically significant predictors at a 0.05 level

Practice Quiz Question 2

ROC curve

Which of the following are true about the receiver operating characteristic (ROC) curve? Check all that apply.

(A) The ROC curve traces the trade-off between the false positive rate and true positive rate of a classifier, depending on the classification threshold

(B) The maximum value for the area under the curve score is 0.5

(C) A random classifier achieves an area under the curve score of 0.5

(D) One way to calibrate the optimal point on the curve is finding the point closest to the upper left-hand corner

Practice Quiz Question 2

ROC curve

Which of the following are true about the receiver operating characteristic (ROC) curve? Check all that apply.

(A) The ROC curve traces the trade-off between the false positive rate and true positive rate of a classifier, depending on the classification threshold

(B) The maximum value for the area under the curve score is 0.5

(C) A random classifier achieves an area under the curve score of 0.5

(D) One way to calibrate the optimal point on the curve is finding the point closest to the upper left-hand corner

Practice Quiz Question 3

Random forests

A random forest is an example of which type of ensemble learning method?

(A) Bagging

(B) Boosting

(C) Voting

(D) Stacking

Practice Quiz Question 3

Random forests

A random forest is an example of which type of ensemble learning method?

(A) Bagging

(B) Boosting

(C) Voting

(D) Stacking

Practice Quiz Question 4

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{i \neq j}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$

Davies-Bouldin

Recall the Davies-Bouldin index, at right. Which of the following are true about the Davies-Bouldin index?

(A) It is used to choose the optimal number of clusters in k-means clustering.

(B) The goal is to maximize the metric.

(C) It takes into account both the distance between clusters and the distance within clusters.

(D) It is monotonically decreasing with the number of clusters.

Practice Quiz Question 4

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{i\neq j}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$

Davies-Bouldin

Recall the Davies-Bouldin index, at right. Which of the following are true about the Davies-Bouldin index?

(A) It is used to choose the optimal number of clusters in k-means clustering.

(B) The goal is to maximize the metric.

(C) It takes into account both the distance between clusters and the distance within clusters.

(D) It is monotonically decreasing with the number of clusters.

Practice Quiz Question 5

Convolutional neural networks

Which of the following is true about pooling layers in convolutional neural networks? Check all that apply.

(A) The most common pooling aggregations are minimum, mean, and maximum

(B) Pooling reduces the dimensionality of the data and network

(C) Pooling helps reduce overfitting

(D) The most common pooling kernel is 2x2 with a stride width of 2

Practice Quiz Question 5

Convolutional neural networks

Which of the following is true about pooling layers in convolutional neural networks? Check all that apply.

(A) The most common pooling aggregations are minimum, mean, and maximum

(B) Pooling reduces the dimensionality of the data and network

(C) Pooling helps reduce overfitting

(D) The most common pooling kernel is 2x2 with a stride width of 2

Practice Quiz Question 6

Decision trees

True or false: A decision tree can learn a nonlinear decision boundary.

(A) True

(B) False

Practice Quiz Question 6

Decision trees

True or false: A decision tree can learn a nonlinear decision boundary.

(A) True

(B) False

Practice Quiz Question 7

Regularization

Which of the following is an example of regularization in a machine learning model? Check all that apply.

(A) Ridge regression

(B) LASSO regression

(C) Decision tree pruning

(D) Dropout layers and sparse neural networks

(E) Principal components analysis

Practice Quiz Question 7

Regularization

Which of the following is an example of regularization in a machine learning model? Check all that apply.

(A) Ridge regression

(B) LASSO regression

(C) Decision tree pruning

(D) Dropout layers and sparse neural networks

(E) Principal components analysis

Practice Quiz Question 8

Regression discontinuity

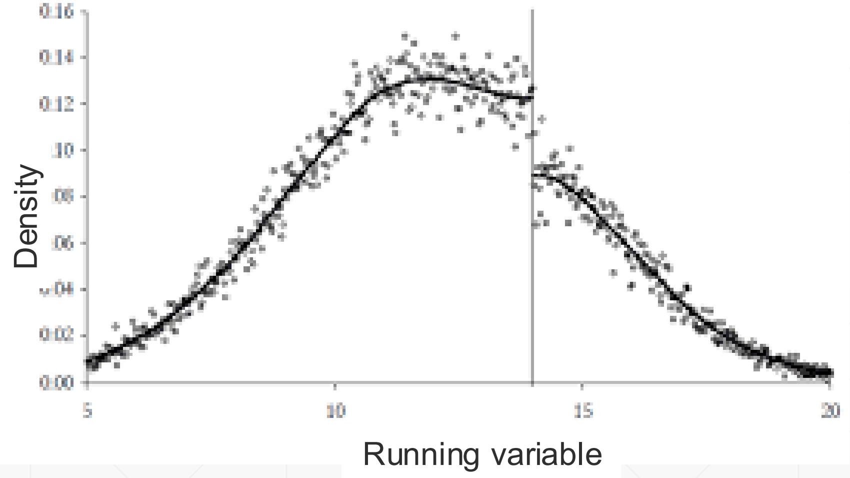
The plot at right of the density of the running variable around a threshold could indicate what for a regression discontinuity design to impact evaluation?

(A) There is visual evidence that the treatment had an impact on the outcome variable.

(B) There is visual evidence that the treatment did not have an impact on the outcome variable.

(C) There is visual evidence that the treatment had an impact on a non-outcome covariate.

(D) There is visual evidence of pre-treatment manipulation of the decision threshold.



Practice Quiz Question 8

Regression discontinuity

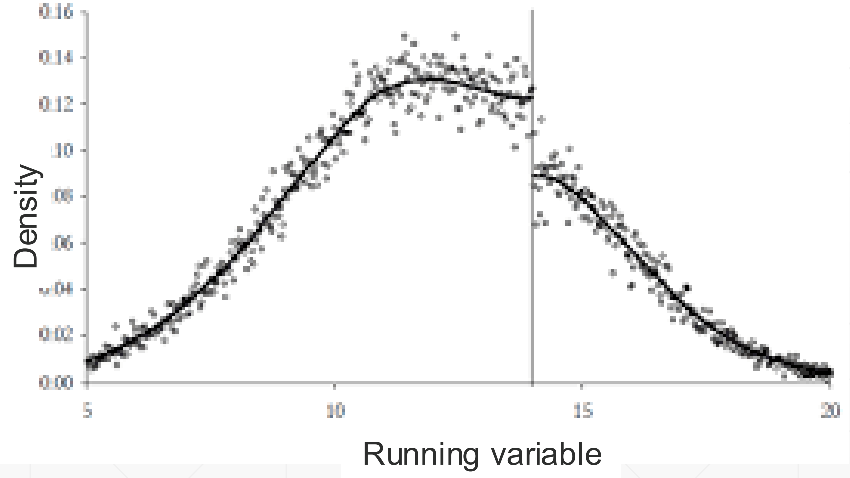
The plot at right of the density of the running variable around a threshold could indicate what for a regression discontinuity design to impact evaluation?

(A) There is visual evidence that the treatment had an impact on the outcome variable.

(B) There is visual evidence that the treatment did not have an impact on the outcome variable.

(C) There is visual evidence that the treatment had an impact on a non-outcome covariate.

(D) There is visual evidence of pre-treatment manipulation of the decision threshold.



Practice Quiz Question 9

Multiclass classification

You are evaluating a classification model for predicting the number in a handwritten digit image from the MNIST dataset. You study examples where the real digit was a 7 but the classifier predicted a 3. This is an example of…

(A) Ablative analysis

(B) Error analysis

(C) Feature importances

(D) SHAP values

Practice Quiz Question 9

Multiclass classification

You are evaluating a classification model for predicting the number in a handwritten digit image from the MNIST dataset. You study examples where the real digit was a 7 but the classifier predicted a 3. This is an example of…

(A) Ablative analysis

(B) Error analysis

(C) Feature importances

(D) SHAP values

Practice Quiz Question 10

Imputation

You are analyzing panel data that tracks poverty over time. You notice that two covariates associated with poverty – education and race – are missing for over 60% of observations in one year of your data. Which would be an appropriate way to deal with the missing data? Select all that would be appropriate.

(A) Drop the observations with missing data

(B) Drop the features with missing data

(C) Model-based imputation, using other covariates to predict education and race

(D) Carry forward education and race from a previous year

(E) Mean, median, or mode imputation of education and race

(F) Zero imputation of education and race

Practice Quiz Question 10

Imputation

You are analyzing panel data that tracks poverty over time. You notice that two covariates associated with poverty – education and race – are missing for over 60% of observations in one year of your data. Which would be an appropriate way to deal with the missing data? Select all that would be appropriate.

(A) Drop the observations with missing data

(B) Drop the features with missing data

(C) Model-based imputation, using other covariates to predict education and race

(D) Carry forward education and race from a previous year

(E) Mean, median, or mode imputation of education and race

(F) Zero imputation of education and race

Practice Quiz Question 11

Principal components analysis

Which of the following are true about principal components analysis (PCA)? Select all that apply.

(A) The principal components are the eigenvectors of the data's correlation matrix.

(B) PCA is deterministic: If run twice on the same dataset for the same number of components *k*, the results will be the same.

(C) The eigenvalues tell you how much variation in the original dataset is explained by each principal component.

(D) The first PCA component for a decomposition with 1 component will be the same as the first PCA component for a decomposition with 10 components.

(E) PCA should be calculated on standardized features.

Practice Quiz Question 11

Principal components analysis

Which of the following are true about principal components analysis (PCA)? Select all that apply.

(A) The principal components are the eigenvectors of the data's correlation matrix.

(B) PCA is deterministic: If run twice on the same dataset for the same number of components *k*, the results will be the same.

(C) The eigenvalues tell you how much variation in the original dataset is explained by each principal component.

(D) The first PCA component for a decomposition with 1 component will be the same as the first PCA component for a decomposition with 10 components.

(E) PCA should be calculated on standardized features.

Practice Quiz Question 12

Feature importance

Which of the following are methods for calculating feature importance in decision trees, random forests, and other tree-based models? Select all that apply.

(A) Calculate the mean weighted decrease in impurity from splitting on a feature

(B) SHAP partial dependence plots

(C) Permutation importance

(D) Count the number of times that a feature is split on in the tree or forest

Practice Quiz Question 12

Feature importance

Which of the following are methods for calculating feature importance in decision trees, random forests, and other tree-based models? Select all that apply.

(A) Calculate the mean weighted decrease in impurity from splitting on a feature

(B) SHAP partial dependence plots

(C) Permutation importance

(D) Count the number of times that a feature is split on in the tree or forest

Practice Quiz Question 13

Spam v/s Ham

Suppose you are building a classifier to separate spam (y = 1) emails from non-spam/ham (y = 0) emails. In your training dataset, 98% of the emails are non-spam/ham, while the remainder are spam. Which of the following are true?

(A) If you always predict spam (y = 1), your classifier has recall 100%, and precision of 2%

(B) If you always predict non-spam (y = 0), your classifier has accuracy 98%

(C) If you always predict spam (y = 1), your classifier has recall 0%, and precision of 98%

(D) If you always predict non-spam (y = 0), your classifier has recall 0%

Practice Quiz Question 13

Spam v/s Ham

Suppose you are building a classifier to separate spam (y = 1) emails from non-spam/ham (y = 0) emails. In your training dataset, 98% of the emails are non-spam/ham, while the remainder are spam. Which of the following are true?

(A) If you always predict spam (y = 1), your classifier has recall 100%, and precision of 2%

(B) If you always predict non-spam (y = 0), your classifier has accuracy 98%

(C) If you always predict spam (y = 1), your classifier has recall 0%, and precision of 98%

(D) If you always predict non-spam (y = 0), your classifier has recall 0%

Practice Quiz Question 14

Word embeddings

You have a vocabulary of size N1, and you decide to generate embeddings of dimension N2 (i.e the embedding for each word **w** *is a vector = [w_1, w_2, …. W_{N2}]. Which of the following is / are true?*

(A) N1 >> N2

(B) The cosine similarity between a pair of word embeddings increases as similarity increases

(C) The Euclidean distance between a pair of word embeddings increases as similarity increases

(D) N1 = N2

Practice Quiz Question 14

Word embeddings

You have a vocabulary of size N1, and you decide to generate embeddings of dimension N2 (i.e the embedding for each word **w** *is a vector = [w_1, w_2, …. W_{N2}]. Which of the following is / are true?*

(A) N1 >> N2

(B) The cosine similarity between a pair of word embeddings increases as similarity increases

(C) The Euclidean distance between a pair of word embeddings increases as similarity increases

(D) N1 = N2

Practice Quiz Question 15

Gradient descent

Which of the following is true about gradient descent? Select all that apply.

(A) After each iteration, we modify the weight vector in the direction of the negative gradient

(B) Each update of the weight vector depends on all the training examples

(C) Gradient descent always converges to the global minimum

(D) After each iteration, we modify the weight vector in the direction of the gradient

(E) If your training dataset is large, stochastic gradient descent is preferable to gradient descent

Practice Quiz Question 15

Gradient descent

Which of the following is true about gradient descent? Select all that apply.

(A) After each iteration, we modify the weight vector in the direction of the negative gradient

(B) Each update of the weight vector depends on all the training examples

(C) Gradient descent always converges to the global minimum

(D) After each iteration, we modify the weight vector in the direction of the gradient

(E) If your training dataset is large, stochastic gradient descent is preferable to gradient descent