INFO251 – Applied Machine Learning

Lab 12

Emily Aiken

Announcements

- **PS7** due Monday May 2
- **Quiz 2** on Thursday, April 28
  - **Let us know via email or piazza if you have a DSP accommodation or time conflict**

Agenda

- Topics covered in AML
- ML algorithms “cheat sheet”
- 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. Optimization and Loss Functions
- Mean squared error
- Logistic loss
- Hinge (RELU) loss
- Cross entropy loss
- Gradient descent

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

5. Neural Networks
- Perceptron
- Fully Connected Networks
- Autoencoders
- Convolutional Neural Networks
- Recurrent Neural Networks

6. Fairness
- Independence, sufficiency, separation
- Protected attributes and privilege
- p% rule
- Thresholding
- Fairness constrained classification

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

8. Practical ML
- Train-test splits
- Cross validation
- Imputation
- Normalization
- Standardization
- Feature engineering
- Imbalanced data
- Regularization
- Overfitting
- Bias-variance trade-off
- Interpretability

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, catboost, lightgbm`	Gradient boosting models
`keras and tensorflow`	Neural networks

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: 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: Neural Networks

Algorithm	Applications	Hyperparameters	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 • Number of fully connected layers at the end • Regularization/dropout	Convolutional layers use matrix multiplication to 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		
Autoencoder	Reconstruction	• Number of nodes in hidden layer (degree of dimensionality reduction) • Activation functions	By training to predict the input, outputs of hidden layer are lower dimensional embedding of input		

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.		

Practice Quiz Question 1

Linear regression

Using the Boston Housing dataset, you run a linear regression to predict the median house value of a neighborhood based on whether it is adjacent to the Charles river (RIV) and the crime rate (CRIM). The results are at right. Which of the following are true?

	Coefficient	95% confidence interval
Intercept	35	[33.6, 37.2]
RIV	9.7	[7.6, 10.8]
CRIM	-1.3	[-3.7, 0.2]

(A) An area far from the Charles with no crime would have an expected median housing value of $35
(B) For a 1% increase in the crime rate, there is a 1.3% decrease in housing value on average
(C) Being next to the Charles river increases housing value by $9.7 on average
(D) Both crime rate and adjacency to the Charles river are 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) One way to calibrate the optimal point on the curve is finding the point closest to the upper left hand corner

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

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

Practice Quiz Question 3

Computational complexity

Rank the following models from least to most expensive computation in the training phase: k nearest neighbors, LASSO regression, naïve bayes, random forest, neural network

(A) LASSO regression < naïve bayes < k nearest neighbors < NN< random forest

(B) Naïve bayes < k nearest neighbors < random forest < LASSO regression < NN

(C) K nearest neighbors < naïve bayes < LASSO regression < random forest < NN

(D) K nearest neighbors < LASSO regression < NN < random forest < naïve bayes

Practice Quiz Question 4

Fairness

Which of the following strategies can help ameliorate bias in machine learning classifiers? Check all that apply.

(A) “Fairness through awareness”

(B) Alternative classification boundaries for protected classes

(C) Leaving protected features out of the training data

(D) Fairness constrained classification

Practice Quiz Question 5

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 6

Clustering

Which of the following are requirements for a clustering distance metric? Check all that apply.

(A) Symmetric

(B) Non-negative

(C) Convex

(D) Satisfies Fisher's inequality

(E) Satisfies triangle inequality

Practice Quiz Question 7

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

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

(C) The goal is to maximize the metric.

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

Practice Quiz Question 8

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 9

Decision trees

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

(A) True

(B) False

Practice Quiz Question 10

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