Example Final Exam

NYU CDS: Introduction to Data Science

May 11, 2019

Problem 1

(a) Why is a random search more efficient than grid search for finding the optimal hyper-parameter values of an estimator in practice?
(b) Describe a method for finding hyper-parameters of an estimator which outperforms both grid search and random search.
(c) Describe five components of a pipeline for machine learning and data science. Give a concrete example for each component. Suggest a method for optimizing the hyper-parameters of such a pipeline.
(d) Give an example of data augmentation for an image dataset and for a tabular dataset. Why may data augmentation improve the pipeline performance?

Problem 2

(a) Define statistical (demographic) parity for group fairness. Define equal opportunity for group fairness. Give an example of using each fairness criteria.
(b) Can an arbitrary classifier ensure group fairness by all the following three criteria simultaneously: false positive rate, false negative rate, and positive predictive value? Describe each criteria and explain.
(c) Give an example where a classification decision has not only an immediate impact on fairness, but also a delayed impact on fairness. Describe a method for measuring that impact.

Problem 3

(a) Describe the difference between content-based recommendation and collaborative filtering. Give an example for using each.
(b) Give pseudocode for an iterative algorithm for collaborative filtering.
(c) Define the singular value decomposition (SVD).
(d) The SVD separates a matrix into rank one pieces in order of importance. Define the first k pieces for rank k.

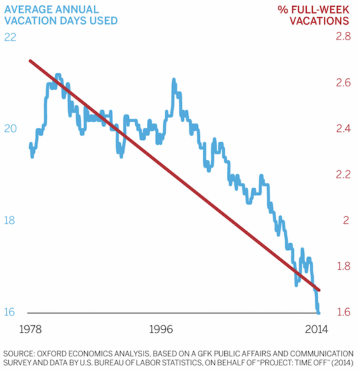
(a) What is the SVD of the matrix $\begin{pmatrix} 2 & 4 \\ 1 & 2 \end{pmatrix}$?

Problem 4

(a) Define prior distribution, posterior distribution, and likelihood function. Describe a method for estimating the posterior distribution of a parameter.
(b) Given a neural network classifying 100 classes you are given information that the test set will be balanced (equal number of examples for each class). How can you use this information to improve the network performance?
(c) Given multiple neural network classifiers, describe two ensemble methods for combining their results into a single better result.

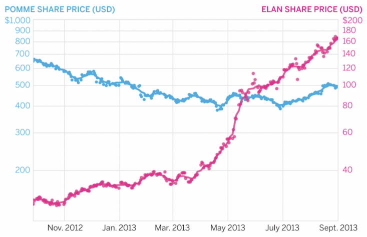
Problem 5

Describe a different problem in each of the following 4 charts, and suggest a way to correct each graph.

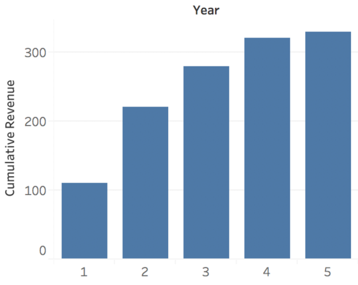


SOURCE: OXFORD ECONOMICS ANALYSIS, BASED ON A GFK PUBLIC AFFAIRS AND COMMUNICATION SURVEY AND DATA BY U.S. BUREAU OF LABOR STATISTICS, ON BEHALF OF "PROJECT: TIME OFF" (2014)

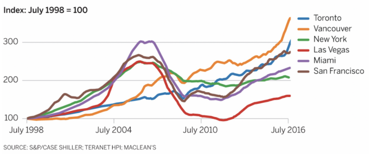
(a)



(b)



(c)



SOURCE: S&P/CASE SHILLER; TERANET HPI; MACLEAN'S

(d)