Assignment 2: Gaussian Mixtures and Linear Discriminant Analysis

Due: April 29, 2022 at 11:59 pm

1 Instructions

The answers to the questions and the code should be submitted on GradeScope by 29 April 2022, 11:59 pm. You may submit the notebook downloaded as PDF, but please make sure that the questions are clearly segmented and labelled. You are encouraged to use the submission template provided. To secure full marks for a question both the answer and the code should be correct. Completely wrong (or missing) code with correct answer will result in zero marks.

2 Data and Task Description

Download the 'hw2_data.csv' data file from Canvas. This data tracks yearly fundamentals of publicly traded companies from the NYSE, NASDAQ, and AMEX. The target variables labeled as 'label' is a categorical variable of gross profitability defined as $\frac{Earnings_t - COGS_t}{TotalAssets_{t-1}}$. The feature space includes 58 variables defined in Tables 1 and 2. This task, in essence, is a forecasting exercise. The feature space **is already lagged**. Therefore, **you do not need to lag the variables yourself**. The data track these companies over 3 years (2018-2020). We will train the data in 2018, validate in 2019, and forecast 2020. Further instructions are given in the questions.

3 Generative Learning

1. (5 points) Keep only the labels between -1 and 3.

   Split the data into Train-Validation-Test:

   - Training data should contain features in 2018, do not forget to remove 'label'
   - Training labels should only contain 'label' in 2018
   - Validation data should contain features in 2019, do not forget to remove 'label'
   - Validation labels should only contain 'label' in 2019
   - Test data should contain features in 2020, do not forget to remove 'label'
   - Test labels should only contain 'label' in 2020

2. (10 points) Compute and report the prior probabilities π_j for all labels in the Training set.

3. (15 points) Using the Training set, calculate the likelihood for feature 'ret' to be 0.1 conditional on each value of the label $P_j = P(ret = 0.1|y = j)$ (Use the Normal PDF found in the following link: Scipy). Report the density for each value of the label. You need to code this by hand, 0 points will be given if you use the pre-coded scipy function.

   Note: You can use the scikit-learn function from Question 4 onwards.

4. (10 points) Use Guassian naive bayes from the scikit-learn library (found here:scikitlearnfunction) to classify the test data. Report the accuracy. You need to use the train+validation set.

5. (10 points) Compute the confusion matrix (as shown in the lectures) and report the top 2 pairs with most (absolute number) incorrect classifications.

6. (15 points) Implement Gaussian Mixture model on the data as shown in class. Tune the covariance type parameter on the validation data. Use the selected value to compute the test accuracy. As always, train the model on train+validation data to compute the test accuracy. Train the model twice, the first model should use the covariance type that yielded the highest accuracy in the validation stage. The second model should use the covariance type that yielded the second highest accuracy in the validation stage. Comment on the accuracy on the test set of the the models you ran. **Hint: Use 'n_components=3, init_params="kmeans", random_state=34'**.

7. (5 points) Bonus Question: Apply Linear Discriminant Analysis model on the train+validation data and report the accuracy obtained on test data. Report the transformation matrix (w) along with the intercept.

4 Appendix

Table 1: Variable Description

Variable	Description
ret	Return on Investment
acc	Operating Accruals
agr	Asset growth
bm	Book-to-market equity
cfp	Cashflow-to-price
ep	Earnings-to-price
ni	Net Stock Issues
op	Operating profitability
rsup	Revenue surprise
cash	Cash holdings
chcsho	Change in shares outstanding
cashdebt	Cash to debt
pctacc	Percent operating accruals
lev	Leverage
sgr	Sales growth
sp	Sales-to-price
invest	Change in Fixed Investments
roe	Return on Equity
lgr	Growth in long-term debt
roa	Return on Asset
depr	Depreciation / PPandE
egr	Equity Growth
chato	2 Yr Mean Sales Growth
chtx	Change in tax expense
noa	(Changes in) Net Operating Assets
rna	Quarterly Return on Net Operating Assets, Quarterly Asset Turnover
pm	profit margin
ato	Asset Turnover

Table 2: Variable Description Continued

Variable	Description
dy	Dividend yield
roic	Return on Investment-Current
chinv	Mean Change in Inventory
pchsale_pchinvt	Percent Change in Sales - Percent Change in Inventory
pchsale_pchrect	Percent Change in Sales - Percent Change in Receivables
pchgm_pchsale	Percent Change in Gross Margin - Percent Change in Sales
pchsale_pchxsga	Percent Change in Sales - Percent Change in SGA
pchdepr	Percent Change in Depr. Amort.
pchcapx	Percent Change in CAPX
grcapx	2 Year CAPX Growth
grGW	Goodwill Growth
currat	Current Asset to Liability Ratio
quick	Quick Ratio
salecash	Sales to Cash Ratio
salerec	Sales to Receivables
saleinv	Sales to Inventory
pchsaleinv	Percent Change in Sales to Inventory
grltnoa	Growth in long-term net operating assets
conv	Deferred Charges to LT Debt
operprof	Operating Profit
capxint	CAPX to 2 Yr Mean Total Asset
chpm	Industry-adjusted change in profit margin
alm	Quarterly Asset Liquidity
me	Market equity
hire	Employee growth rate
herf	Industry sales concentration
bm_ia	Industry-adjusted book to market
gp	Gross Profit
me_ia	Industry-adjusted size