Introduction to Machine Learning

Problem Set 3

Q1. Data Clean-Up

a. (One-Hot Encoding) Report in a new table the resulting dataset after encoding the categorical features with a one-hot encoding.

Dataset	Type_Comedy	Type_Drama	Length_Short	Length_Medium	Length_Long	IMDB	Liked?
Train	0	1	1	0	0	6.8	Yes
Train	1	0	1	0	0	9.1	Yes
Train	1	0	0	0	1	5.2	No
Train	1	0	0	1	0	6.8	No
Train	0	1	0	1	0	8.1	Yes
Train	1	0	1	0	0	4.7	No
Train	0	1	0	1	0	8.1	Yes
Train	0	1	1	0	0	7.6	Yes
Train	1	0	0	0	1	3.5	Yes
Test	0	1	0	0	1	8.2	Yes

So,
Number of features before One-Hot Encoding → 3
Number of features after One-Hot Encoding → 6

b. (Imputing Missing Values) Report in a new table the resulting dataset after imputing all missing values using the feature mean for examples belonging to the same class.

Dataset	Type_Comedy	Type_Drama	Length_Short	Length_Medium	Length_Long	IMDB	Liked?
Train	0	1	1	0	0	6.8	Yes
Train	1	0	1	0	0	9.1	Yes
Train	1	0	0	0	1	5.2	No
Train	1	0	0	1	0	6.8	No
Train	0	1	0	1	0	8.1	Yes
Train	1	0	1	0	0	4.7	No
Train	0	1	0	1	0	8.1	Yes
Train	5/7	2/7	0	1	0	6.7	Yes
Train	0	1	1	0	0	7.6	Yes
Train	1	0	0	0	1	3.5	Yes
Train	1	0	1/7	3/7	3/7	7.1	Yes
Test	1	0	1	0	0	8.2	Yes
Test	0	1	0	0	1	7.1	Yes
Test	5/7	2/7	0	1	0	7.4	Yes

1. Missing value for categorical features -
   For deciding the values for categorical variables, I have calculated mean of all the observations belonging to the same class i.e. either class – Yes or class – No.

2. Missing value for quantitative feature –
   I have calculated mean of all the training observations present in the same class. i.e. either class – Yes or class – No.

c. (Feature Scaling) Report in a single table the IMDb values for all data as follows:
   Column 1: resulting values after min-max scaling
   Column 2: resulting values after standardization

Original (IMDb Rating)	Min-Max Scaling (IMDb Rating)	Standardization (IMDb Rating)
6.8	0.58	0
9.1	1	1.5
5.2	0.3	-1
6.8	0.58	0
8.1	0.82	0.86
4.7	0.21	-1.4
8.1	0.82	0.86
7.6	0.73	0.53
6.7	0.57	-0.06
3.5	0	-2.2
7.1	0.64	0.2
8.2	0.83	0.93
7.1	0.64	0.2
7.4	0.69	0.4

Mean value → 6.8
Standard Deviation → 1.5

Q2. Dimensionality Reduction (5 points)

a. Name three uses for dimensionality reduction techniques.

1. Overcome the causes of high dimensional data – Dimensionality reduction is mainly used to avoid risks/ problems associated with high dimensional data. There can be a dataset with high number of predictor variables but a smaller number of observations. So, in such cases there is a high risk of overfitting. To avoid this, it makes sense to reduce the dimensions without losing any kind of information. So, dimensionality reduction is important.
2. Efficient computations – As we are reducing number of variables, specially the variables which are less significant in terms of their impact on predictor/ output variable. It helps in reducing use of memory as well as computation power and time.
3. Better visualization – Reducing dimensions/ variables makes picture of a system clearer.
4. No loss of data/ information – With all above advantages, there is no risk of loss of any meaningful information associated with reducing number of variables.

b. In your own words, describe when to use Principle Component Analysis (PCA) versus Locally Linear Embedding (LLE) to reduce the feature dimensionality.

1. Principle Component Analysis (PCA) is a dimensionality reduction technique which transforms the high dimensional data to a fewer dimensions in a linear fashion. In contrast, Locally Linear Embedding (LLE) is also a dimensionality reduction technique which transforms the high dimensional data to a fewer dimensions in a non-linear fashion.
2. For dimensionality reduction, PCA works out a linear mapping of the given data where as LLE works on manifold technique.
3. PCA fits n-dimensional ellipsoid/ principle components to the given data, LLE finds given number of neighbors of each point and then calculates weights for each point.
4. For PCA, length of an axis of an ellipsoid determines the variance. For ex. Smaller the length of the axis, there is a small amount of variance present along that axis. For LLE, after computing set of weights/ linear combinations of neighbors for each point, apply optimization techniques to find low-dimensional space with information about all the linear combination of neighbors.

Q3. Classification Evaluation (4 points)

a. What is a precision-recall curve" (PR curve) and how can it be used when designing a machine learning system?

Precision and Recall are inversely proportional. That is, whenever Precision increases, recall decreases and vice-versa. So, in order to build a system, we need to maintain/ achieve a right balance between Precision and Recall. Precision-Recall curve helps in achieving this balance and to compare the performance. In standard classification problem, precision is the classified instances that are correctly classified, while recall is the fraction of classified instances to all correctly classified instances. To understand and decide the trade-off between the precision and recall, precision-recall curve can be used. An area near to the boundary under the curve defines high precision and a high recall.

b. What is the relationship between a PR curve and a confusion matrix?

Confusion matrices don't do anything on their own. One can create different metrics of evaluating a built model using different values stored in a confusion matrix. So, we can say confusion matrices are not meaningful/ useful for evaluations, but they can be used to create and understand different evaluation metrics.

Formulae for confusion matrix, precision, recall –

Confusion matrix –

	Predicted (Yes)	Predicted (No)
Actual (Yes)	TP	FN
Actual (No)	FP	TN

Precision = TP/ (TP + FP)

Recall = TP / (TP + FN)

Here,
TP – True Positive
FP – False Positive
FN – False Negative
TN – True Negative

c. What is a ROC curve" and how can it be used when designing a machine learning system?

ROC stands for Receiver Operating Characteristics. ROC curve is a tool use for evaluating and comparing different models in predictive modeling. In machine learning, ROC curve defines clarity on how different statistical models differentiate between 'True Positive' and 'True Negative' values. The ROC curve plots Sensitivity vs Specificity. Here, Sensitivity concerned about probability of predicting true positive as a positive and on the other hand, Specificity concerned about probability of predicting true negative as a positive. The optimistic situation for any model will be the curve which is high on sensitivity and low on specificity.

d. What is the relationship between a ROC curve and a confusion matrix?

Same as for PR curve (Q3. C), confusion matrices don't do anything on their own. One can create different metrics of evaluating a built model using different values stored in a confusion matrix. So, we can say confusion matrices are not meaningful/ useful for evaluations, but they can be used to create and understand different evaluation metrics.

ROC is a continuous measure unlike confusion matrix. Confusion matrix is used when you have clear distinct categories and now you want to understand the behavior. ROC curve describes range of possible values of true positives and false positives.

Q4. Challenge Analysis (8 points): find two machine learning competitions (challenges) on the platform Kaggle. For each competition, write a response to the following items:

Challenge 1: https://www.kaggle.com/c/new-york-city-taxi-fare-prediction/data

a. Describe the motivation for the competition.
The challenge was hosted by Google cloud and Coursera to predict the fare amount (inclusive of tolls) for a taxi ride in New York city based on pickup and drop off locations. Before this, if cab drivers estimate the fare based on only the distance between the two points then there was a difference of $5-$8 been recorded. So, task was to improve the prediction system using different machine learning techniques.

b. Describe the machine learning task. You must include a discussion of what is the raw input (e.g., images, text, etc) and what are the target labels a machine learning must predict.

The goal of the given challenge is to predict 'fare_amount' for each trip using number of predictor variables. The raw input is in 'text' format. More precisely input data consist of numeric/ quantitative data. And as I mentioned earlier, there is only one target variable i.e. 'fare_amount' to predict the taxi fare. So, in this case machine learning task is to build a statistical model that will handle and analyze all the given quantitative predictor variables and make a prediction about a taxi fare amount in New York city.

c. Describe how the dataset is partitioned for training and evaluation (e.g., train/val/test split?) and how many examples are included for each partition.

Dataset is divided into two parts. Two different csv files are provided for the analysis. One file contains input features and target fare_amount values for all the observations present in training set. And one file containing test set observations for which values for predictor variables have been given and task is to predict the taxi fare amount i.e. value of a target variable.

Number of observations in training dataset → around 55M rows.

Number of observations in testing dataset → around 10K rows.

Number of input predictor variables → 6

Number of target variables → 1

d. Describe the evaluation metrics used to assess the performance of algorithms submitted for the competition.
RMSE (root mean squared error) is the evaluation metric used in this competition. So, RMSE basically measures the difference between the predicted taxi fare amount value and

the corresponding actual taxi fare value. As it is related to error, smaller the value of RMSE it's a better model to predict.

Challenge 2: https://www.kaggle.com/c/restaurant-revenue-prediction#evaluation

a. Describe the motivation for the competition.
The challenge is hosted by TFI an investment company who is largely investing in developing new restaurant sites. Deciding a location for a restaurant is a very important decision as restaurants sites requires lot of money and time investments to run and sustain their business. So, goal is to find a statistical model to do the location level analysis before investing in a new restaurant site.

b. Describe the machine learning task. You must include a discussion of what is the raw input (e.g., images, text, etc) and what are the target labels a machine learning must predict.
The goal of this challenge is to predict revenue of the restaurant in a given year using different machine learning algorithms. The raw input in this case is in 'text' format. More precisely input data consist of numeric/ quantitative and categorical data. And as I mentioned earlier, there is only one target variable i.e. 'revenue' to predict the financial progress of a restaurant. So, in this case machine learning task is to build a statistical model that will handle and analyze all the given quantitative and converted qualitative predictor variables and make a prediction about a revenue amount of a newly open restaurant in a given year.

c. Describe how the dataset is partitioned for training and evaluation (e.g., train/val/test split?) and how many examples are included for each partition.

Dataset is divided into two parts. Two different csv files are provided for the analysis. One file contains input features and target revenue values for all the observations present in training set. And one file containing test set observations for which values for predictor variables have been given and task is to find the revenue i.e. value of a target variable.

Number of observations in training dataset → 137 rows.

Number of observations in testing dataset → 100K i.e. 100000 rows.

Number of input predictor variables → 42

Number of target variables → 1

d. Describe the evaluation metrics used to assess the performance of algorithms submitted for the competition.

   RMSE (root mean squared error) is the evaluation metric used in this competition. So, RMSE basically measures the difference between the predicted revenue value and the corresponding actual revenue value. As it is related to error, smaller the value of RMSE it’s a better model to predict.