CS – 344 Guide 9 – Classification

1. Google's Machine Learning Crash Course
    a. Logistic Regression
        i. Definition: a model that generates a probability for each possible discrete label value in classification problems by applying a sigmoid function to a linear prediction.
        ii. Terms:
            - *Sigmoid*
                1. *A function that maps logistic or multinomial regression output (log odds) to probabilities, returning a value between 0 and 1.*
                2. *Y = 1 / (1 + e^(-alpha))*
                    1. *Alpha = b + w1x1 + w2x2 + …wnxn (logistic regression)*
            - *Log loss*
                1. *The loss function used in binary logistic regression*
                2. *= Riemann sum (x,y) element of D (-y*log(y')-(1-y)*log(1-y')*
            - *Entropy*
                1. *(cross-entropy) a generalization of Log Loss to multi-class classification problems.*
                2. *Quantifies the difference between two probability distributions.*
            - *Likelihood function*
                1. *A particular function of the parameter of a statistical model given data.*
                2. *Gives an idea of how well the data summarizes unknown parameters in probability distributions.*
        iii. Compare and contrast *logistic* vs. *linear* regression.
            - Logistic regression values are in the range of (0, 1).
                1. Used when the dependent variable is binary in nature. (or categorical – not quantitative in nature)
            - Linear regression values are in the range of [0,1]
                1. Used when the dependent variable is continuous in nature
                2. Used when the regression line is linear in nature.
    b. Classification
        i. Terms:
            - *ROC curve (receiver operating characteristic curve)*
                1. *A curve of true positive rate versus false positive rate at different classification thresholds*
                2. *Area under the ROC curve is the probability that a classifier will be more confident that a randomly chosen positive example is actually positive than that a randomly chosen negative example is positive.*
            - *Prediction bias*
                1. *A value indicating how far apart the average of predictions is from the average of labels in the dataset.*
            - *Calibration plot (calibration layer)*
                1. *A post-prediction adjustment, typically to account for prediction bias.*

2. *The adjusted predictions and probabilities should match the distribution of an observed set of labels*

ii. Compare and contrast:
- *regression* vs. *classification*.
  1. *Regression model: a type of model that outputs continuous (typically, floating-point) values.*
  2. *Classification model: a type of model for distinguishing among two or more discrete classes.*
- *accuracy* vs. *precision* vs. *recall*.
  1. *Accuracy: the fraction of predictions that a classification model got right*
     1. *Multi-class: accuracy = correct predictions / total # of examples*
     2. *Binary-class: accuracy = true positives + true negatives / total # of examples*
  2. *Precision: a metric for classification models that identifies the frequency with which a model was correct when predicting the positive class.*
     1. *Precision = true positives / true positives + false positives*
  3. *Recall: a metric for classification models that answers the following question:*
     1. *Out of all the possible positive labels, how many did the model correctly identify?*
     2. *Recall = true positives / true positives + false negatives*

c. Regularization for Sparsity

i. Terms:
- *Convex optimization*
  1. *The process of using mathematical techniques such as gradient descent to find the minimum of a convex function.*

2. Google's ML Practicum: Image Classification — Study the first two sections: "Introduction" – "Check Your Understanding".

i. Why doesn't simple network like the one we used for the MNIST dataset work in general?

i. The MNIST dataset uses very small images where all neurons in the network are connected to each other. This is not very scalable and manageable when you move up to larger images in the 1080p-2160p (1920x1080 to 3840x2160) range. At this point you would have, image length x image width x 3 (R,G,B) neurons connected to each other in just the first layer, let alone more layers. Therefore, it would be too computationally expensive to do it this way. Hence, using convolutional neural networks for image classification is the way to go.

ii. Terms

i. *Convolution*
- *Extracts tiles of the input feature map, and applies filters to them to compute new features, producing an output feature map, or convolved feature.*
- *Mixes the convolutional filter and the input matrix in order to train weights*
- *Refers to either convolutional operation or convolutional layer.*

1. *Defined by two parameters:*
    1. *Size of the tiles that are extracted (typically 3x3 or 5x5 pixels)*
    2. *Depth of the output feature map, corresponding to the # of filters that are applied.*

ii. *Convolved Feature*
- An output feature map produced by a convolution operation.
- Can have different size and depth than the input feature map.

iii. *Pooling*
- *Reducing a matrix (or matrices) created by an earlier convolutional layer to a smaller matrix.*
- *Usually involves taking either maximum or average value across the pooled area.*
- *Often called subsampling or downsampling*
- *Helps enforce translational invariance in the input matrix.*