Ali Mirzazadeh

HW1, CS4641

- For each of the following, state whether you would use binary classification, multi-class classification, or regression and give a one or two sentence justification.
    - Given a mobile phone customer's account age, monthly bill, usage data and other factors, predict the likelihood that the customer will switch to another provider at the end of their current contract.

      I would use binary classification because the prediction is whether a person will switch to another provider, and that is a binary output. Therefore, classifying based on age, monthly bill, usage data, and other factors to determine this boolean output would be a binary classification

    - Given ages, household income, and other factors, predict the number of children a couple will have.

      The output is non-negative integers up to the maximum number of children, so it not continuous but rather in multiple bins, up to say, 10. The program would have to decide which bin to place datapoint in based on the ages, household income, and other factors so this would be a type of multi-class classification.

    - Given a traveler's answers to a series of questions, decide whether the the traveler is allowed across the bridge or not.

      The output is a binary variable that states whether a traveler should be allowed to cross a bridge or not (true can cross, false cannot). Using the questions, the program would have to classify the traveler into one of these two cases so it is a binary classification.

    - Given a college graduate's undergraduate major, institution, GPA, major GPA and other factors, predict their salary.

      The output is a continuous positive number with a large range of possible values, so using the major, institution, GPA, major GPA, and other factors I would create a regression to figure out the salary.

- What are the general steps of most learning algorithms?

Start by assigning each feature a weight of 1. Then, going through the training data set one by one, adjust the weights of the features to minimize the in-sample error, but be careful to not adjust the weights too much and overfit the data. Once finished with the training set, the set of weights is the optimal parameters.

- What is the difference between online learning algorithms and bath learning algorithms?

Online learning algorithms constantly update the weights with each datapoint and every error calculation. Batch learning algorithms, however, use the average of many input data points to update the weights so it is not necessarily the best optimization at each individual point but over time it reaches the best point. Batch learning is therefore cheaper to calculate for larger datasets, whereas online is used for smaller ones.

- How can binary classifiers be used to solve a multiclass classification problem?

Multiple binary classifiers could serve as inputs, which would be assigned weights that would, when multiplied by the input binary classifiers, sum to any bin in the multiclass output.

- How can linear models deal with data sets that are not linearly separable due to noise in the data?

Linear models, instead of terminating when the in sample error is 0, would be assigned a threshold of acceptable in sample error rate. This means that whenever the in sample error is below this threshold, the program would terminate. This would allow for noise in the data without causing an infinite loop.

- What is the form of the hypothesis class of linear classifiers?

The hypothesis class of linear classifiers is a weight vector that corresponds to a linear classification

- What are the functional forms and the loss functions for
    - **the perceptron learning algorithm**
      Functional Form:

      $$\mathcal{H} = \{h(\vec{x}) = sign(\vec{w}^T \cdot \vec{x})\}$$

      Loss Function:

$$E(h) = \frac{1}{N}\sum_{n=1}^{N}[\![h(\vec{x}) \neq f(\vec{x})]\!]$$

- **linear regression**
  Functional Form:

  $$\nabla E_{in}(\vec{w}) = \frac{2}{N}(X^T X\vec{w} - X^T\vec{y})$$

  Loss Function:

  $$E_{in} = \frac{1}{N}\sum_{n=1}^{N}(h(\vec{x_n} - y_n)^2$$

- **logistic regression.**
  Functional Form:

  $$h(\mathbf{x}) = \theta\left(\sum_{i=0}^{d} w_i x_i\right) = \theta(\mathbf{w}^{\mathrm{T}}\mathbf{x})$$

  Loss Function:

  $$E_{\mathrm{in}}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}\ln(1 + e^{-y_n \cdot \mathbf{w}^{\mathrm{T}}\mathbf{x}_n})$$

- What is a local minimum?

  A location on a surface that has the lowest value within the area selected but not necessarily the lowest value over the entire function (the global minimum)

- What is a global minimum?

  A point on a surface corresponding to the minimum value in the entire surface.

- What is the meaning of the learning rate hyperparameter to a learning algorithm?

  The learning rate hyperparameter sets how quickly the weight vector adopts each datapoint in the training set and adjusts to its error.

- What may happen if the learning rate is set too low?

The weight change would be very small, taking a long time, and would also be prone to falling into local minima.

- What may happen if the learning rate is set too high?

  The weight change would be very large, and could possibly miss the minimum because it jumps over it since the change is very large.

- Describe two desirable features of the sigmoid loss function for logistic regression.
  1. **The sigmoid loss function is differentiable**
  2. **Provides a value equal to the probability of event occurrence [0,1]**
- What is a feature transform?

  A feature transform is applying a function to a certain feature to transform it so that it is easier to distinguish or differentiate the data points, and is often used to make the data linearly separable.

- Can a linear model be used to separate into classes the feature vectors of instances that are not linearly separable? If so, how?

  Yes. By performing a feature transform on the training data, we could make the data linearly separable and make what once were non-linear classifications into linear ones.

- What happens to sample complexity (the number of training samples we need to maintain a bound our generalization error) with higher-order polynomial transforms?

  The sample complexity increases exponentially, so it becomes very important to limit the order of the transform polynomial as much as possible. The simpler the polynomial, the less data is requried.