Homework 1

Aleck Zhao

February 22, 2018

1) **Supervised vs. Unsupervised Learning**

1. Give an example of a problem that could be solved with both supervised and unsupervised learning. Is data readily available for this problem? How would you measure your 'success' in solving this problem for each approach?

   Answer. Image analysis. Supervised learning could be trained on image data, while unsupervised learning could use clustering to identify objects. Data would be readily available for this problem. Supervised learning would be successful if objects are correctly identified, and unsupervised learning would be successful if objects could be identified.

2. What are the pros and cons of each approach? Which approach do you think the problem better lends itself to?

   Answer. Pros of supervised: It will probably perform better can be tailored to the dataset.

   Cons of supervised: The model might be overtrained, and it is computationally intensive.

   Pros of unsupervised: Can discover patterns that a labeled dataset might not include. Simpler to implement.

   Cons of unsupervised: Parameters must be set beforehand.

   Supervised learning would be better for this application if we want the learning algorithm to be able to identify a certain image, such as faces.

2) **Model Complexity** Explain when you would want to use a simple model over a complex model and vice versa. Are there any approaches you could use to mitigate the disadvantages of using a complex model?

Answer. A simple model is preferred when the hypothesis class is simple and is expected to have relatively few factors, and a complex model is preferred when the hypothesis class could have many factors. We can use cross-validation to mitigate the disadvantages of using a complex model to avoid over-fitting.

3) **Training and Generalization** Suppose you're building a system to classify images of food into two categories: either the image contains a hot dog or it does not. You're given a dataset of 25,000 (image, label) pairs for training and a separate dataset of 5,000 (image, label) pairs.

1. Suppose you train an algorithm and obtain 96% accuracy on the larger training set. Do you expect the trained model to obtain similar performance if used on newly acquired data? Why or why not?

   Answer. I expect the trained model to obtain similar performance on new data because the training set was very large, so the trained model should achieve the correct result very often. Because the training set was large, we can approximate the true function very well.

2. Suppose that, after training, someone gives you a new test set of 1,000 (image, label) pairs. Which do you expect to give greater accuracy on the test set: The model after trained on the dataset of 25,000 pairs or the model after trained on the dataset of 5,000 pairs? Explain your reasoning.

   Answer. The model after training on the 25,000 dataset should give greater accuracy for the reasons stated above.

3. Suppose your models obtained greater than 90% accuracy on the test set. How might you proceed in hope of improving accuracy further?

   Answer. If we train the model on more and diverse data, we will probably obtain an even greater accuracy.

4) **Loss Function** State whether each of the following is a valid loss function for binary classification. Wherever a loss function is not valid, state why. Here, y is the correct label and $\hat{y}$ is a decision confidence value, meaning that the predicted value is given by $\text{sign}(\hat{y})$ and the confidence on the classification increases with $|\hat{y}|$.

1. $\ell(y, \hat{y}) = \frac{3}{4}(y - \hat{y})^2$

   Answer. This is valid. The loss is always non-negative, and there exists a predictive function $\hat{y}$ that makes the loss 0 ($\hat{y} = y$).

2. $\ell(y, \hat{y}) = |(y - \hat{y})| / \hat{y}$

   Answer. This is not valid. The loss can be arbitrarily large if $\hat{y}_i = -c$ for $c > 0$.

3. $\ell(y, \hat{y}) = \max(0, 1 - y \cdot \hat{y})$.

   Answer. This is valid. The loss is always non-negative since it is at least 0, and there exists a predictive function $\hat{y}$ that makes the loss 0 ($\hat{y} = 1/y$)

5) **Linear Regression** Suppose you observe n data points $(x_1, y_2), \cdots, (x_n, y_n)$, where all x_i and all y_i are scalars.

1. Suppose you choose the model $\hat{y} = wx$ and aim to minimize the sum of squares error $\sum_i (y_i - \hat{y}_i)^2$. Derive the closed form solution for w from scratch, where 'from scratch' means without using the least-squares solution presented in class.

   Solution. The sum of squares error is given by

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 &= \sum_{i=1}^{n}(y_i - wx_i)^2 = \sum_{i=1}^{n}(y_i^2 - 2wx_iy_i + w^2x_i^2) \\ &= \sum_{i=1}^{n} y_i^2 - 2w\sum_{i=1}^{n} x_iy_i + w^2\sum_{i=1}^{n} x_i^2\end{aligned}$$

   To minimize this, take the partial with respect to w and setting equal to 0,

$$\begin{aligned}0 &= \frac{\partial}{\partial w}\left(\sum_{i=1}^{n} y_i^2 - 2w\sum_{i=1}^{n} x_iy_i + w^2\sum_{i=1}^{n} x_i^2\right) = -2\sum_{i=1}^{n} x_iy_i + 2w\sum_{i=1}^{n} x_i^2 \\ \implies w &= \frac{\sum_{i=1} x_iy_i}{\sum_{i=1} x_i^2}\end{aligned}$$

   □

2. Suppose you instead choose the model $\hat{y} = w \sin x$ and aim to minimize the sum of squares error $\sum_i (y_i - \hat{y}_i)^2$. Is there a closed-form solution for w? If so, what is it?

   Solution. The sum of squares error is given by

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 &= \sum_{i=1}^{n}(y_i - w\sin x_i)^2 = \sum_{i=1}^{n}(y_i^2 - 2wy_i\sin x_i + w^2\sin^2 x_i) \\ &= \sum_{i=1}^{n} y_i^2 - 2w\sum_{i=1}^{n} y_i\sin x_i + w^2\sum_{i=1}^{n}\sin^2 x_i\end{aligned}$$

   To minimize this, take the partial with respect to w and setting equal to 0,

$$\begin{aligned}0 &= \frac{\partial}{\partial w}\left(\sum_{i=1}^{n} y_i^2 - 2w\sum_{i=1}^{n} y_i\sin x_i + w^2\sum_{i=1}^{n}\sin^2 x_i\right) = -2\sum_{i=1}^{n} y_i\sin x_i + 2w\sum_{i=1}^{n}\sin^2 x_i \\ \implies w &= \frac{\sum_i y_i\sin x_i}{\sum_i \sin^2 x_i}\end{aligned}$$

   □

6) **Logistic Regression** Explain whether each statement is true or false. If false, explain why.

1. In the case of binary classification, optimizing the logistic loss is equivalent to minimizing the sum-of-squares error between our predicted probabilities for class 1, $\hat{\mathbf{y}}$, and the observed probabilities for class 1, $\mathbf{y}$.

   Answer. This is false. The sum-of-squares error function is not convex, whereas logistic loss is convex, so these are not equivalent.

2. One possible advantage of stochastic gradient descent is that it can sometimes escape local minima. However, in the case of logistic regression, the global minimum is the only minimum, and stochastic gradient descent is therefore never useful.

   Answer. This is false. The fact that the global minimum is the only minimum means that stochastic gradient descent will always converge to the correct minimum.