VEF ACADEMY

MACHINE LEARNING 2018

Homework 3

November 28, 2018

- This homework is due at 2 PM, December 8, 2018.
- Please submit the HW via Google Form (Link will be sent out shortly). Code for programming problems should be submitted as .py files.
- You can discuss HW problems with the instructor, TAs, classmates, or others, but the work you submit must be your own work.

- You may write your answers in Vietnamese or English or a mix of both languages.
- You may consult textbooks and print and online materials.
- Please show all of your work. Answers without appropriate justification will receive very little credit. For programming questions, please submit all the code.

Problem 1. *(10 points)* In the "New York City Taxi Fare Prediction" Competition on Kaggle (https://www.kaggle.com/c/new-york-city-taxi-fare-prediction/data), competitors have to predict the cost (in USD) of a taxi ride in New York City given the following features:

- *pickup_datetime* – timestamp value indicating when the taxi ride started.
- *pickup_longitude* – float for longitude coordinate of where the taxi ride started.
- *pickup_latitude* – float for latitude coordinate of where the taxi ride started.
- *dropoff_longitude* – float for longitude coordinate of where the taxi ride ended.
- *dropoff_latitude* – float for latitude coordinate of where the taxi ride ended.
- *passenger_count* – integer indicating the number of passengers in the taxi ride.

Competitors are given a training dataset (*train.csv*) with input features and target *fare_amount* values. They will then have to predict the *fare_amount* for each row of input features in a test set (*test.csv*). Using Tom M. Mitchell's definition of Machine Learning discussed in class (for Parts (a) and (b)),
(a) Describe the experience E and the class of tasks T of the algorithms used to solve this problem.
(b) Propose a performance measure P that we can use to rank the competitors' submissions.
(c) Is this problem a supervised learning one or an unsupervised learning one? Justify your answer.
(d) Is this problem a classification or a regression problem? Justify your answer.

Problem 2. *(20 points)* In a homework at the Machine Learning class, Toan uses Logistic Regression to classify customers of a Consumer Finance Company (CFC) into two categories: Low-risk (Negative) and High-risk (Positive). Comparing the output of his model with the loan performance data of 1000 customers,Toan ends up with the following confusion matrix (For a description, see, for example https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/):

n=1000	Predicted Low-risk	Predicted High-risk
Actual Low-risk	850	50
Actual High-risk	20	80

(a) Calculate True Positive rate, False Positive rate, True Negative rate, and False Negative rate.
(b) Discuss the costs (to the CFC) of a False Positive and a False Negative.
(c) Calculate the Accuracy, Precision, Recall, and F_1 score of this classifier.

For references, see also

- https://en.wikipedia.org/wiki/Precision_and_recall,
- https://en.wikipedia.org/wiki/F1_score.

Problem 3 *(35 points)* To prevent overfitting in Linear Regression, we can use a technique called regularization in which we add a penalty term in the loss function to discourage higher values of the coefficients $w_j, j = 1, \ldots, D$. The loss function discussed in class will then become

$$L(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{N} \left(y^{(i)} - \mathbf{w}^T x^{(i)} \right)^2 + \frac{\lambda}{2} \sum_{j=1}^{D} w_j^2 \qquad (1)$$

where we use the constant λ to adjust the effect of the regularization term. Calculate the gradient $\nabla_{\mathbf{w}} L(\mathbf{w})$.

Problem 4 *(35 points)* In Regularized Logistic Regression, the loss function can be written as

$$L(\mathbf{w}) = -\frac{1}{N}\sum_{i=1}^{N}\left[y^{(i)}log(\sigma(\mathbf{w}^T\mathbf{x}^{(i)})) + (1-y^{(i)})log(1-\sigma(\mathbf{w}^T\mathbf{x}^{(i)}))\right] + \frac{\lambda}{2N}\sum_{j=1}^{D} w_j^2$$

where we use the constant λ to adjust the effect of the regularization term. Calculate the gradient $\nabla_{\mathbf{w}} L(\mathbf{w})$.