EE 660 # Homework Week 3 Posted: 9/5/2018

Due: Tue., 9/11/2018, 2:00 PM

Logistic Regression

Reading

i. *Introduction*: Murphy 1.4.6

ii. *Introduction and explanation*. AML p. 88 (starting with Sec. 3.3) - p. 96, up to the block diagram in Ex. 3.4. If you have had gradient descent before, you can skip that part.

Also, *please note* that $E_{in}(\underline{w})$ in Eq. (3.9) and afterwards, refers to the "in sample" error, or error on the training set data. This serves the same purpose as the objective function $J(\underline{w}, \mathcal{D})$ that we have been working with in the past.

iii. *Methods.* Murphy, 8.1, 8.2, 8.3.1-8.3.2, 8.3.6. (It's less reading than it looks like here.)

Problems (AML book)

1. (a) AML Exercise 3.5 (a) (p. 90)
   (b) Find values for $\theta(s)$ and for $\tanh(s)$, for: $s \to +\infty$, $s = 0$, $s \to -\infty$.
   (c) Verify that $1-\theta(s) = \theta(-s)$.
2. (a) Fill in the steps from the equation before (3.9), to Eq. (3.9), on p. 91.
   (b) For the error measure of Eq. (3.9):
      (i) Let n be one data point; for $\underline{y}_n = +1$, what $\underline{w}^T \underline{x}_n$ will yield a large contribution to the error?
      (ii) Let m be a different data point; for $\underline{y}_m = -1$, what $\underline{w}^T \underline{x}_m$ will yield a large contribution to the error?
      (iii) Consider a 2-class classifier in which the discriminant function is $g(\underline{x}) = \underline{w}^T \underline{x}$. For $\underline{y}_n = +1$, compare the size of the contribution to the error (call it $E_n^{(c)}$) for data point $\underline{x}_n$ being correctly classified, with it being incorrectly classified (call it $E_n^{(inc)}$); which is larger? Justify your answer by showing your reasoning.

Homework 3 continues on next page…

Convexity review

Read Murphy 7.3.3 (also refer to Discussions 2 and 3)

3. Convexity and minimization of quadratic functions.
   (a) You are given that $f(\underline{w}) = (\underline{a}^T \underline{w} - b)^2$, in which $\underline{a}$ and $\underline{w}$ are D dimensional vectors, and $\underline{a}$ and b are given constants. Prove that f is convex.

   (b) Is $J(\underline{w}) = \left\| \underline{\underline{X}} \cdot \underline{w} - \underline{y} \right\|_2^2 + \underline{c}^T \underline{w} = \sum_{i=1}^{N} \left(\underline{x_i}^T \underline{w} - y_i \right)^2 + \underline{c}^T \underline{w}$ convex, in which $\underline{x_i}$ and $\underline{w}$ are D dimensional, and $\underline{x_i}$ and $\underline{y_i}$, $i = 1, \ldots, N$, and $\underline{c}$, are given constants? Justify your answer.

Feasibility and fundamental issues of learning

Reading

AML 1.3 (p.15) to end of Ch. 1 (p. 32). Note: if you are short on time, you may skip Section 1.4; we won't need it right away, but you will be responsible for the material later.

Comments on notation and terminology in AML:

- "Sample" means a set of data points or a set of marbles. We can also think of our training dataset as being a "sample".
- $f(\underline{x})$ is the "target function", and denotes the true function that gives the correct output (class label) for an input $\underline{x}$. This function is typically unknown to us. We try to find some reasonable approximation to f by learning from the training data.

Problem

4. Suppose our "learning algorithm" uses a standard linear model for $\hat{f}$ in a classification problem, in which there are D input variables (features):

$$\hat{f}(\underline{x}) = \text{sgn}\left(\underline{w}^T \underline{x}\right)$$

The learning algorithm picks the best weight vector $\underline{\hat{w}}$ using the training data $\mathcal{D}$, based on minimizing some objective function $J(\underline{w}, \mathcal{D})$, with each component of $\underline{w}$ restricted to:

$$w_0 = 1; \quad w_j \in \{1, 2\} \quad \forall j \in \{1, 2, \cdots, D\}.$$

   (a) How many elements (hypotheses) are there in the hypothesis set $\mathcal{H}$?
   (b) How would the Hoeffding Inequality be applied to this case? That is, give an expression, if possible, for an upper bound on $P\left[\left| E_{in}(\hat{h}) - E_{out}(\hat{h}) \right| > \varepsilon \right]$.