MATH 156: Machine Learning — Project #2

October 6, 2020

This is due on October 22 by 10:59 am (PST). You should submit a PDF with neatly written solutions to the mathematical problems, and a short description of your work for the programming problems (in particular highlight what your code does and what data you used). Please attach your code to the end of the PDF.

1. Answer the following questions about kernels.
   - Suppose $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ are disjoint subsets of the components in $\boldsymbol{x}$ such that every component of $\boldsymbol{x}$, x_i, is in either in $\boldsymbol{x}_a$ or $\boldsymbol{x}_b$. Suppose $k_a(\boldsymbol{x}_a, \boldsymbol{x}_a')$ and $k_b(\boldsymbol{x}_b, \boldsymbol{x}_b')$ are valid kernels over their respective spaces. Show that $k(\boldsymbol{x}, \boldsymbol{x}') = k_a(\boldsymbol{x}_a, \boldsymbol{x}_a') + k_b(\boldsymbol{x}_b, \boldsymbol{x}_b')$ is a valid kernel.
   - For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^2$, $k(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^\top \boldsymbol{y})^2$ is a kernel. What is a nonlinear feature space mapping $\boldsymbol{\phi}$ which produces this kernel (i.e., $k(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\phi}(\boldsymbol{y})$)?

2. Recall the polynomial regression model which fits data with functions of the form $y(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + ... + w_M x^M$. Assume that the target t associated to x is distributed as $t \sim \mathcal{N}(y(x, \boldsymbol{w}), \beta^{-1})$, and that there is i.i.d. training data $x_1, x_2, ..., x_N$ with targets $t_1, t_2, ..., t_N$. Show that the maximum likelihood estimator for $\boldsymbol{w}$ is given by the solution to the set of linear equations

   $$\sum_{j=0}^{M} A_{ij} w_j = T_i$$

   where $A_{ij} = \sum_{n=1}^{N} (x_n)^{(i+j)}$ and $T_i = \sum_{n=1}^{N} (x_n)^i t_n$.

3. Show that $\mathbf{u} := A(A^T A)^{-1} A^T \mathbf{v}$ is the orthogonal projection of $\mathbf{v}$ onto the space spanned by the columns of A.

4. Suppose that we are performing linear regression for sampled data points $(\mathbf{x}_1, t_1), (\mathbf{x}_2, t_2), ..., (\mathbf{x}_N, t_N)$ and that while sampling, along with the data point $(\mathbf{x}_i, t_i)$, we are given the scalar value $r_i > 0$ which represents how certain we are about the target value t_i associated to $\mathbf{x}_i$. We can alter the least-squares error function to use this information as

   $$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} r_n \{t_n - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n)\}^2.$$

   Show that the solution $\mathbf{w}^*$ which minimizes this error function is

   $$\mathbf{w}^* = \left(\sum_{n=1}^{N} r_n \boldsymbol{\phi}(\mathbf{x}_n) \boldsymbol{\phi}(\mathbf{x}_n)^T \right)^{-1} \left(\sum_{n=1}^{N} r_n \boldsymbol{\phi}(\mathbf{x}_n) t_n \right).$$

   Hint: You may use the fact that $\frac{\partial}{\partial \boldsymbol{y}} (a - \boldsymbol{y}^\top \cdot \boldsymbol{c})^2 = -2 \cdot (a - \boldsymbol{y}^\top \cdot \boldsymbol{c}) \cdot \boldsymbol{c}$.

5. At `https://archive.ics.uci.edu/ml/datasets.php?format=&task=reg&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=table`, download the ".csv" file for the red wines in the "Wine Quality" dataset. You will now write a program/notebook which loads the data from this csv file, performs linear regression using the closed-form solution derived in class, performs the least-mean-squares (LMS) algorithm, and plots errors between the iterates of the LMS algorithm and the closed-form linear regression solution. After you have loaded the data into the data matrix X and the vector of targets $\boldsymbol{t}$, answer perform the tasks described below and answer the related questions. For this dataset, the targets are the wine quality scores and the input data consists of the other 11 features.
   - Compute $\boldsymbol{w}^*$ as given in the notes. Print the first 5 entries of $\boldsymbol{w}$ and measure $\|X^\top \boldsymbol{w}^* - \boldsymbol{t}\|^2/N$. What do you think of this average error (i.e., do you think the trained regression model is good)?
   - Implement the LMS algorithm for linear regression with stepsize $\eta = 1/\|\boldsymbol{\phi}(\boldsymbol{x}_n)\|^2$ as given in (1) where n is sampled from $1, 2, ..., N$ uniformly at random in each iteration. Let the initial iterate be $\boldsymbol{w}^{(0)} = \mathbf{0}$. Run 100000 iterations of the algorithm and create a plot of $\|\boldsymbol{w}^* - \boldsymbol{w}^{(k)}\|$ for $k = 0, 1, ..., 100000$. Measure $\|X^\top \boldsymbol{w}^{(100000)} - \boldsymbol{t}\|^2/N$.
   - The LMS method described above defines

     $$\boldsymbol{w}^{(k+1)} = \boldsymbol{w}^{(k)} + (t_n - (\boldsymbol{w}^{(k)})^\top \boldsymbol{x}_n)\boldsymbol{x}_n / \|\boldsymbol{x}_n\|^2. \tag{1}$$

     Show that $(\boldsymbol{w}^{(k+1)})^\top \boldsymbol{x}_n = t_n$ where n is the same index as in (1).