Assignment 2

Foundations of Machine Learning
IIT-Hyderabad
Aug-Dec 2021

Max Marks: 30
Due: 6th Oct 2021 11:59 pm

This homework is intended to cover theory and programming exercises in the following topics:

- SVM, Kernels

Instructions

- Please upload your submission on Google Classroom by the deadline mentioned above. Your submission should comprise of a single file (PDF/ZIP), named `<Your_Roll_No>_Assign2`, with all your solutions.
- For late submissions, 10% is deducted for each day (including weekend) late after an assignment is due. Note that each student begins the course with 7 grace days for late submission of assignments (of which atmost 4 can be used for a given submission). Late submissions will automatically use your grace days balance, if you have any left. You can see your balance on the FoML Marks and Grace Days document (will be shared on Piazza).
- Please use PYTHON for the programming questions.
- Please read the department plagiarism policy. Do not engage in any form of cheating - strict penalties will be imposed for both givers and takers. Please talk to instructor or TA if you have concerns.

Questions: Theory

1. **Support Vector Machines:** ***(4 marks)*** In the derivation for the Support Vector Machine, we assumed that the margin boundaries are given by $\mathbf{w}.\mathbf{x}+b=+1$ and $\mathbf{w}.\mathbf{x}+b=-1$. Show that, if the +1 and -1 on the right-hand side were replaced by some arbitrary constants $+\gamma$ and $-\gamma$ where $\gamma > 0$, the solution for the maximum margin hyperplane is unchanged. (You can show this for the hard-margin SVM without any slack variables.)
2. **Support Vector Machines:** ***(4 marks)*** Consider the half-margin of maximum-margin SVM defined by ρ, i.e. $\rho = \frac{1}{||\mathbf{w}||}$. Show that ρ is given by:

$$\frac{1}{\rho^2} = \sum_{i=1}^{N} \alpha_i$$

where α_i are the Lagrange multipliers given by the SVM dual (as on Slide 30 of the SVM lecture uploaded on Piazza). *(Hint: The answer involves just 3-4 steps, if you are thinking of something longer, re-think!)*

3. **Kernels:** ***(5 marks)*** Let k_1 and k_2 be valid kernel functions. Comment about the validity of the following kernel functions, and justify your answer with proof or counter-examples as required:

   (a) $k(x, z) = k_1(x, z) + k_2(x, z)$

   (b) $k(x, z) = k_1(x, z)k_2(x, z)$

   (c) $k(x, z) = h(k_1(x, z))$ where h is a polynomial function with positive co-efficients

   (d) $k(x, z) = \exp(k_1(x, z))$

   (e) $k(x, z) = \exp\left(\frac{-\|\mathbf{x}-\mathbf{z}\|_2^2}{\sigma^2}\right)$

Questions: Programming

4. **SVMs:** ***(2 + 2 + 4 + 2 = 10 marks)*** In this question, you will be working on a soft-margin SVM. You may find it helpful to review the Scikit Learn's SVM documentation: `http://scikit-learn.org/stable/modules/svm.html`.

   We will apply soft-margin SVM to handwritten digits from the processed US Postal Service Zip Code data set. The data (extracted features of intensity and symmetry) for training and testing are available at:

   - http://www.amlbook.com/data/zip/features.train
   - http://www.amlbook.com/data/zip/features.test

   In this dataset, the 1st column is digit label and 2nd and 3rd columns are the features. We will train a one-versus-one (one digit is class +1 and another digit is class -1) classifier for the digits '1' (+1) and '5' (-1). (In the original dataset, only consider data samples(rows) with the label as either 1 or 5, for both train and test settings. Then for training details, you may find this link at http://scikit-learn.org/stable/modules/svm.html helpful.)

   (a) Consider the linear kernel $K(\mathbf{x}_n, \mathbf{x}_m) = \mathbf{x}_n^T\mathbf{x}_m$. Train using the provided training data and test using the provided test data, and report your accuracy over the entire test set, and the number of support vectors.

   (b) In continuation, train only using the first $\{50, 100, 200, 800\}$ points with the linear kernel. Report the accuracy over the entire test set, and the number of support vectors in each of these cases.

   (c) Consider the polynomial kernel $K(\mathbf{x}_n, \mathbf{x}_m) = (1 + \mathbf{x}_n^T\mathbf{x}_m)^Q$, where Q is the degree of the polynomial. Comparing $Q = 2$ with $Q = 5$, comment whether each of the following statements is TRUE or FALSE.

      i. When $C = 0.0001$, training error is higher at $Q = 5$.

      ii. When $C = 0.001$, the number of support vectors is lower at $Q = 5$.

      iii. When $C = 0.01$, training error is higher at $Q = 5$.

iv. When $C = 1$, test error is lower at $Q = 5$.

(d) Consider the radial basis function (RBF) kernel $K(\mathbf{x}_n, \mathbf{x}_m) = e(-||\mathbf{x}_n - \mathbf{x}_m||^2)$ in the soft-margin SVM approach. Which value of $C \in \{0.01, 1, 100, 10^4, 10^6\}$ results in the lowest training error? The lowest test error? Show the error values for all the C values.

Deliverables:

- Code, Brief report (PDF) with your solutions for the above questions

5. **SVMs (contd):** ***(3 + 4 = 7 marks)*** GISETTE (`https://archive.ics.uci.edu/ml/datasets/Gisette`) is a handwritten digit recognition problem. The problem is to separate the highly confusible digits '4' and '9'. This dataset is one of five datasets of the NIPS 2003 feature selection challenge. The dataset for this problem is large, so please budget time accordingly for this problem.

(a) *Standard run:* Use all the 6000 training samples from the training set to train the model, and test over all test instances, using the linear kernel. Report the train error, test error, and number of support vectors.

(b) *Kernel variations:* In addition to the basic linear kernel, investigate two other standard kernels: RBF (a.k.a. Gaussian kernel; set $\gamma = 0.001$), Polynomial kernel (set `degree` = 2, `coef0` = 1; e.g, $(1 + \mathbf{x}^T\mathbf{x})^2$). Which kernel yields the lowest training error? Report the train error, test error, and number of support vectors for both these kernels.

Deliverables:

- Code, Brief report (PDF) with your solutions for the above questions