Stony Brook University
CSE512 – Machine Learning – Spring 17
Homework 3, Due: March, 28, 2017, 11:59PM

Instructions

- The homework is due on March 28, 2017. Anything that is received after the deadline will not be considered.
- The write-up **must** be prepared in Latex, including the Matlab code and figures in the report, and converted to pdf.
- We can use any Latex class you like, just report question number and your answer.
- If the question requires you to implement a Matlab function, the answer should be your code. Make sure it is sufficiently well documented that the TAs can understand what is happening.
- Each Question, regardless of how many sub-questions contains, is worth 10 points.

Preliminaries

Suppose labeled points $(\boldsymbol{x}, y)$ are drawn from $\mathcal{X} \times \mathbb{R}$, where $\mathcal{X}$ is the feature space. Suppose you are also given a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which computes the inner product for a feature map $\phi : \mathcal{X} \to \mathcal{H}$ where $\mathcal{H}$ is a Hilbert space for k. In other words,

$$k(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle \ .$$

Suppose you have collected a training set of m examples

$$S = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_m, y_m) \in \mathcal{X} \times \mathbb{R}\} \ .$$

In the following, everytime we use an algorithm with a kernel k, we mean using the algorithm on the transformed dataset

$$S' = \{(\phi(\boldsymbol{x}_1), y_1), (\phi(\boldsymbol{x}_2), y_2), \cdots, (\phi(\boldsymbol{x}_m), y_m) \in \mathcal{H} \times \mathbb{R}\} \ .$$

A Gaussian kernel is defined as $k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\gamma \|\boldsymbol{x} - \boldsymbol{x}'\|_2^2\right)$. For all the experiments below, set $\gamma = 0.001$.

For the linear kernel, simply $\phi(\boldsymbol{x}) = \boldsymbol{x}$, and $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{x}, \boldsymbol{x}' \rangle$.

The provvided dataset is a subset of the 'Adult' dataset: the task is to predict if a person makes over 50K a year, see `https://archive.ics.uci.edu/ml/datasets/Adult`.

1 Support Vector Machines

In this problem, you will implement the soft-margin SVMs without bias b using two different optimization techniques: (1) quadratic programming and (2) gradient descent.

1.1 Question 1: Primal and Dual of Kernel SVM

Quadratic programs refer to optimization problems in which the objective function is quadratic and the constraints are linear. Quadratic programs are well studied in optimization literature, and there are many efficient solvers. Many Machine Learning algorithms are reduced to solving quadratic programs, as you already saw in the previous assignments. In this question, we will use the quadratic program solver of Matlab to optimize the dual objective of a kernel SVM.

The primal objective of a kernel SVM without the bias b can be written as

$$\min_{\boldsymbol{w}\in\mathcal{H}} \ \frac{1}{2}\left\|\boldsymbol{w}\right\|_2^2 + C\sum_{j=1}^{m}\ell(\boldsymbol{w},\phi(\boldsymbol{x}_j),y_j) \ . \tag{1}$$

Here $\ell(\boldsymbol{w},\boldsymbol{x}_j,y_j)$ is the *Hinge loss* of the j-th instance:

$$\ell(\boldsymbol{w},\boldsymbol{x}_j,y_j) = \max\left(1-y_j\langle\boldsymbol{w},\phi(\boldsymbol{x}_j)\rangle,0\right) \ .$$

The corresponding dual objective is

$$\max_{\boldsymbol{\alpha}} \ \sum_{j=1}^{m}\alpha_j - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y_i\alpha_i y_j\alpha_j k(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{2}$$

$$0\le\alpha_j\le C \ \forall j \ . \tag{3}$$

(a) Write the SVM dual objective as a quadratic program. Look at the `quadprog` function of Matlab, and write down what `H, f, A, b, Aeq, beq, lb, ub` are.

(b) From the Lagrangian of the above problem, show that $\boldsymbol{w} = \sum_{i=1}^{m}\alpha_i y_i\phi(\boldsymbol{x}_i)$.

1.2 Question 2: Implement a kernel SVM using Quadratic Programming

(a) Using the answer to part (a) in Question 1, use quadratic programming to write a function that solves the dual SVM objective. In Matlab, you can use the function `quadprog`. The prototype must be

```
alpha = train_ksvm_dual(X, y, C, kernel, gamma)
```

where `kernel` is 'gaussian' for Gaussian kernels and 'linear' for linear kernels, `gamma` is the parameter of the Gaussian kernel.

(b) Using the answer to part (b) in Question 1, write a function to test the solution obtained above, producing a list of predictions from the inputs $\boldsymbol{\alpha}$, the training data, and the test samples. The prototype must be

```
ypredicted = test_ksvm_dual(alpha, Xtr, ytr, Xte, kernel, gamma)
```

where `kernel` is 'gaussian' for Gaussian kernels and 'linear' for linear kernels, `gamma` is the parameter of the Gaussian kernel.

(c) Set $C = 10$, train and test your SVM implementation with linear and Gaussian kernel on the provvided dataset and report the accuracy, the objective value, and the number of support vectors.

(d) Repeat the above question with $C = .1$.

1.3 Question 3: Implement Linear SVM using Sub-Gradient Descent

In class we saw that Sub-Gradient Descent can be used to optimize a convex and not differentiable objective function. Here, we will use it to optimize the SVM primal objective in (1) for the linear kernel.

We can use sub-gradient descent to optimize this objective. Denote by $L(\boldsymbol{w})$ the function $\frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{j=1}^{m}\ell(\boldsymbol{w}, \boldsymbol{x}_j, y_j)$. The update rule for $\boldsymbol{w}$ at iteration t is

$$\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \eta_t \partial_{\boldsymbol{w}} L(\boldsymbol{w}_t), \tag{4}$$

where $\partial_{\boldsymbol{w}} L(\boldsymbol{w}_t)$ denotes the sub-gradient of L w.r.t. $\boldsymbol{w}$ evaluated in $\boldsymbol{w}_t$. The pseudo-code is in Algorithm 1.

Algorithm 1 Sub-gradient descent for linear SVM

$\boldsymbol{w}_1 = \mathbf{0}$
for $t = 1, 2, \cdots, T$ **do**
  $\eta_t \leftarrow \frac{a}{t}$
  Update $\boldsymbol{w}$ using Eq. (4)
end for

(a) Write the sub-gradient descent update rule for $\boldsymbol{w}$ for linear SVMs.

(b) Implement Algorithm 1 for linear SVMs. a is a tunable parameter. The prototype must be

```
[w] = train_svm_sgd(X, y, C, a, T)
```

(c) The theory in [1] tells us that for this optimization problem the optimal (worst-case) setting of the learning rate is $\eta_t = \frac{1}{t}$, that is setting $a = 1$, but if you are free to choose another learning rate if you want. Using the provvided dataset as your training set, run $T = 10000$ iterations using $C = 10$. Be patient: it might take a while to compute.

(d) Plot the value of the objective function in Eq. (1) after each iterations in a log-log plot, to better show the behaviour. Compare with the objective value obtained in Question 2.

(e) Plot the training error after each iteration, on a normal plot, and compare it with the results in Question 2.

(f) Plot the test error after each iteration, on a normal plot, and compare it with the results in Question 2.

(g) Change C to .1 and repeat what you did in the previous four points.

1.4 Question 4: Invariance to Additive Constants in Kernels

Consider the soft-margin kernel SVM formulation, that includes the bias term b:

$$\min_{\boldsymbol{w},b} \ \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{j=1}^{m}\max(1 - y_j(\langle \boldsymbol{w}, \phi(\boldsymbol{x}_j)\rangle + b), 0) \ .$$

Its dual formulation is

$$\max_{\boldsymbol{\alpha}} \ \sum_{j=1}^{m}\alpha_j - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y_i\alpha_i y_j\alpha_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{5}$$

$$\sum_{j}^{m} y_j\alpha_j = 0 \tag{6}$$

$$0 \le \alpha_j \le C \ \forall j \ . \tag{7}$$

In this question we will investigate the effect of adding a constant to the kernel function. In particular, prove that optimal value of both objective functions and the optimal solution $\boldsymbol{w}^*$ does not change if we add a constant c to the kernel, i.e. $k(\cdot,\cdot) \to k(\cdot,\cdot) + c$.

Hint: Starts from the dual formulation and see what happens when the kernel changes as described. The proof is short...

2 Kernel Ridge Regression

Consider solving the ridge regression problem with a training set

$$S = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_m, y_m) \in \mathbb{R}^d \times \mathbb{R}\} \ .$$

Let $\boldsymbol{w}_\lambda$ be the ridge regression solution with regularization parameter λ, i.e.

$$\boldsymbol{w}_\lambda = \underset{\boldsymbol{w}\in\mathbb{R}^d}{\arg\min} \ \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2 + \frac{1}{m}\|\boldsymbol{y} - X\boldsymbol{w}\|_2^2$$

In class and in HW2, we derived a closed form expression for $\boldsymbol{w}_\lambda$ in terms of the input matrix X and label vector $\boldsymbol{y}$, defined as

$$X = \begin{bmatrix} — & \boldsymbol{x}_1 & — \\ — & \boldsymbol{x}_2 & — \\ & \vdots & \\ — & \boldsymbol{x}_m & — \end{bmatrix} \in \mathbb{R}^{m\times d} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^m \ .$$

This closed form expression turns out to be

$$\boldsymbol{w}_\lambda = (X^\top X + m\lambda I)^{-1} X^\top \boldsymbol{y} \ .$$

We will derive an alternative form for $\boldsymbol{w}_\lambda$ without using the duality nor the Representer Theorem.

2.1 Question 5: Alternative Formulation for the Solution of Ridge Regression

First, note that the above closed form expression comes from solving the optimality equation for $\boldsymbol{w}_\lambda$, i.e.

$$(X^\top X + m\lambda I_d)\boldsymbol{w}_\lambda = X^\top \boldsymbol{y},$$

where I_d is the identity matrix of size $d \times d$. The above equation can be rewritten as

$$m\lambda \boldsymbol{w}_\lambda = X^\top(\boldsymbol{y} - X\boldsymbol{w}_\lambda) \ . \tag{8}$$

Now, define $\boldsymbol{\alpha} := \boldsymbol{y} - X\boldsymbol{w}_\lambda$.

(a) Using equation (8) and the definition of $\boldsymbol{\alpha}$, show that $\boldsymbol{\alpha}$ satisfies the following equation:

$$\frac{1}{m\lambda} XX^\top \boldsymbol{\alpha} + \boldsymbol{\alpha} = \boldsymbol{y} \ . \tag{9}$$

(b) Solve equation (9) for $\boldsymbol{\alpha}$ and find an explicit expression of $\boldsymbol{\alpha}$ as a function of $\boldsymbol{y}, X, m$, and λ.

(c) Use the expression of $\boldsymbol{\alpha}$ at the previous point to show that the following alternative expression for $\boldsymbol{w}_\lambda$ is valid:

$$\boldsymbol{w}_\lambda = X^\top (XX^\top + m\lambda I_m)^{-1} \boldsymbol{y} \ . \tag{10}$$

2.2 Question 6: Implement Kernel Ridge Regression

Let $\boldsymbol{w}_\lambda$ be the solution of the ridge regression problem with regularization parameter λ when we use a kernel k corresponding to a feature mapping ϕ.

(a) Implement a function that finds $\boldsymbol{\alpha}$ for the Kernel Ridge Regression formulation with kernels, using the expression found in the previous question. The prototype must be

```
alpha = train_krr(X, y, lambda, kernel, gamma)
```

where `kernel` is 'gaussian' for Gaussian kernels and 'linear' for linear kernels, `gamma` is the parameter of the Gaussian kernel.

Hint: Observe that the matrix $XX^\top$ can be computed using the kernel function k.

(b) Given a matrix of test points $Xte \in \mathbb{R}^{m_{te} \times d}$, use equation (10) to implement a Matlab function that computes the prediction using the ridge regression solution expressed in terms of $\boldsymbol{\alpha}$. The prototype must be

```
ypredicted = test_krr(alpha, Xtr, ytr, Xte, lambda, kernel, gamma)
```

where `kernel` is 'gaussian' for Gaussian kernels and 'linear' for linear kernels, `gamma` is the parameter of the Gaussian kernel.

(c) Train and test your implementations of Kernel Ridge Regression with the provvided dataset with $\lambda = 2e-05$, Gaussian and linear kernel. Report training and test error.

(d) Repeat the above with $\lambda = 0.002$.

References

[1] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal Estimated sub-GrAdient SOlver for SVM. In *Proc. of ICML*, pages 807–814, 2007.