Homework Set #2 : Perceptron, Regression and Kernel

Due 6th May 2018, Sunday, before 11:59 pm

Submission

- Submit your solutions electronically on the course Gradescope site as PDF files.
- If you plan to typeset your solutions, please use the LaTeX solution template. If you must submit scanned handwritten solutions, please use a black pen on blank white paper and a high-quality scanner app.

Problem 1 (Perceptron [15 pts])

Suppose we have a training set with 8 samples, each sample has feature vector in $\mathbb{R}^2$:

#	1	2	3	4	5	6	7	8
X	[4,0]	[1,1]	[0,1]	[-2,-2]	[-2,1]	[1,0]	[5,2]	[3,0]
y	1	-1	-1	1	-1	1	-1	-1

We are going to implement perceptron algorithm to train a linear classifier with 2 dimensional weight vector $\boldsymbol{w} \in \mathbb{R}^2$ (without bias term). We start with initial weight vector as the first sample in our dataset, i.e. $\boldsymbol{w}_1 = \boldsymbol{x}_1$. Note that: when $\boldsymbol{w}^T x = 0$, the algorithm predicts +1.

To simplify the calculation, you only need to test and possibly update each sample once in the given sequence. You can either implement the algorithm by hand or programming.

(a) **(2 pts)** Is the data linearly separable? Will our algorithm converge if we run it several times over the same sequence? Explain.

(b) **(6 pts)** Regardless of whether the dataset is linearly separable or not, calculate the updates of the weight vector on this sequence for one round over the entire dataset. Show your computations.

(c) **(3 pts)** Up to current training stage, provide closed form functions for the perceptron, Voted perceptron and Average perceptron, using the weight vector(s) you derived in b).

(d) **(4 pts)** Using the functions you derived in c), Compare the training errors between perceptron, the Voted perceptron predictor and the Average perceptron predictor after a single run of the algorithm over the entire dataset.

Problem 2 (Logistic Regression [15 pts])

(a) **(4 pts)** Plot the sigmoid function $\sigma(wx) = 1/(1+e^{-wx})$ v.s. $x \in \mathbb{R}$ with weight $w = 1, 5$ and 100. Use these plots to argue why a solution with large weights can cause logistic regression to overfit.
Note: 1) A qualitative sketch is enough. 2) If you want to plot the function by coding, you don't need to submit the source code.

(b) **(6 pts)** Consider the modified objective function that we minimize in regularized logistic regression:

$$J(\boldsymbol{w}) = -\sum_{n=1}^{N} \left[y_n \log h_{\boldsymbol{w}}(\boldsymbol{x}_n) + (1-y_n) \log\left(1 - h_{\boldsymbol{w}}(\boldsymbol{x}_n)\right) \right] + \frac{1}{2} \sum_i w_i^2$$

where $h_{\boldsymbol{w}}(\boldsymbol{x}) = \sigma(\boldsymbol{w}^T x)$ and the sigmoid function $\sigma(\boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{x})}$. Find the partial derivatives $\frac{\partial J}{\partial w_j}$ and derive the gradient descent update rules for the weights.

(c) **(5 pts)** In class you have seen the probabilistic interpretation of the logistic regression objective, and the derivation of the gradient descent rule for maximizing the conditional likelihood M(C)LE.

$$\max_{w_0,\ldots,w_d} \prod_{i=1}^{n} P(y_i|x_i, w_0, \ldots, w_d),$$

To prevent overfitting, we want the weights to be small. To achieve this, we can consider maximum conditional a posterior M(C)AP estimation:

$$\max_{w_0,\ldots,w_d} \prod_{i=1}^{n} P(y_i|x_i, w_0, \ldots, w_d) f(w_0, \ldots, w_d)$$

where $f(w_0, \ldots, w_d)$ is a prior on the weights. The prior corresponds to having a belief about the possible values that $\boldsymbol{w}$ could have taken.

Assume a standard Gaussian prior $\mathcal{N}(0, \boldsymbol{I})$ for the weight vector, and show that the M(C)AP estimation is equivalent to minimizing the modified logistic regression objective given above. Note: For $\boldsymbol{Z} \sim \mathcal{N}(0, \boldsymbol{I}_m)$,

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \frac{1}{(2\pi)^{\frac{m}{2}}} \exp\left(-\sum_i \frac{w_i^2}{2}\right)$$

Problem 3 (Locally Weighted Linear Regression [10 pts])

Consider a linear regression problem in which we want to "weight" different training instances differently because some of the instances are more important than others. Specifically, suppose we want to minimize

$$J(w_0, w_1) = \sum_{n=1}^{N} \alpha_n \left(w_0 + w_1 x_{n,1} - y_n\right)^2. \tag{1}$$

Here $\alpha_n > 0$. In class, we worked out what happens for the case where all the weights (the α_n's) are the same. In this problem, we will generalize some of those ideas to the weighted setting.

(a) **(4 pts)** Calculate the gradient by computing the partial derivatives of J with respect to each of the parameters (w_0, w_1).

(b) **(6 pts)** Prove that Eq. (1) has a global optimal solution.

Problem 4 (Non-linearity of Kernel [15 pts])

Suppose you are given 6 one-dimensional points: 3 with negative labels $x_1 = -1, x_2 = 0, x_3 = 1$ and 3 with positive labels $x_4 = -3, x_5 = -2, x_6 = 3$. In this question, we will compare classification performance of linear classifiers with or without kernel. Note that, for the entire problem, we define an "optimal" parameter of a linear classifier in terms of classification accuracy, i.e. #correctly classified samples/#samples.

(a) **(3 pts)** Consider a linear classifier with weight vector $\boldsymbol{w} \in \mathbb{R}^2$, i.e. $f(x) = \text{sign}(w_1 x + w_0)$. Write down the optimal value of $\boldsymbol{w}$ and its classification accuracy on the above 6 points. There might be more than one optimal solution, writing down one of them is enough.

(b) **(3 pts)** Given two samples x and z in $\mathbb{R}$, define the kernel $K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ as

$$K(x, z) = xz(1 + xz) \tag{2}$$

Find the corresponding feature map $\phi(x)$.

(c) **(3 pts)** Now consider a separating hyperplane as a line, which can be parameterized by its normal equation, i.e. $w_1 y_1 + w_2 y_2 + w_0 = 0$ for appropriate choices of $\boldsymbol{w} \in \mathbb{R}^3$. Here $(y_1, y_2) = \phi(x)$ is the result of applying the feature map $\phi(x)$ you derived in b) to the original feature x. And the classifier is defined in the form of $f(x) = sign(w_1 y_1 + w_2 y_2 + w_0)$. Provide the optimal value of $\boldsymbol{w}$ and its classification accuracy on the above 6 points. There might be more than one optimal solutions, write down one of them is enough.

(d) **(3 pts)** Apply $\phi(x)$ you derived from b) to the data and plot the points in the new $\mathbb{R}^2$ feature space. On the plot of the transformed points, plot the separating hyperplane you found in c).

(e) **(3 pts)** Draw the decision boundary of the separating hyperplane you found in c) in the original $\mathbb{R}^1$ feature space.

Problem 5 (Programming Exercise: Polynomial Regression [45 pts])

In this exercise, you will work through linear and polynomial regression. Our data consists of inputs $x_n \in \mathbb{R}$ and outputs $y_n \in \mathbb{R}, n \in \{1, \ldots, N\}$, which are related through a target function $y = f(x)$. Your goal is to learn a linear predictor $h_{\boldsymbol{w}}(x)$ that best approximates $f(x)$. But this time, rather than using `scikit-learn`, we will further open the "black-box", and you will implement the regression model!

---

code and data

- code : `regression.py`
- data : `regression_train.csv`, `regression_test.csv`

This is likely the first time that many of you are working with `numpy` and matrix operations within a programming environment. For the uninitiated, you may find it useful to work through a `numpy` tutorial first.[1] Here are some things to keep in mind as you complete this problem:

- If you are seeing many errors at runtime, inspect your matrix operations to make sure that you are adding and multiplying matrices of compatible dimensions. Printing the dimensions of variables with the `X.shape` command will help you debug.

[1] Try out `SciPy`'s tutorial (http://wiki.scipy.org/Tentative_NumPy_Tutorial), or use your favorite search engine to find an alternative. Those familiar with Matlab may find the "Numpy for Matlab Users" documentation (http://wiki.scipy.org/NumPy_for_Matlab_Users) more helpful.

- When working with `numpy` arrays, remember that `numpy` interprets the `*` operator as element-wise multiplication. This is a common source of size incompatibility errors. If you want matrix multiplication, you need to use the `dot` function in Python. For example, `A*B` does element-wise multiplication while `dot(A,B)` does a matrix multiply.
- Be careful when handling `numpy` vectors (rank-1 arrays): the vector shapes $1 \times N$, $N \times 1$, and N are all different things. For these dimensions, we follow the the conventions of `scikit-learn`'s `LinearRegression` class[2]. Most importantly, unless otherwise indicated (in the code documentation), both column and row vectors are rank-1 arrays of shape N, not rank-2 arrays of shape $N \times 1$ or shape $1 \times N$.

Visualization [5 pts]

As we learned last week, it is often useful to understand the data through visualizations. For this data set, you can use a scatter plot to visualize the data since it has only two properties to plot (x and y).

(a) **(5 pts)** Visualize the training and test data using the `plot_data(...)` function. What do you observe? For example, can you make an educated guess on the effectiveness of linear regression in predicting the data?

Linear Regression [20 pts]

Recall that linear regression attempts to minimize the objective function

$$J(\boldsymbol{w}) = \sum_{n=1}^{N} (h_{\boldsymbol{w}}(\boldsymbol{x}_n) - y_n)^2.$$

In this problem, we will use the matrix-vector form where

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \qquad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_N^T \end{pmatrix}, \qquad \boldsymbol{w} = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_D \end{pmatrix}$$

and each instance $\boldsymbol{x}_n = \big(1, x_{n,1}, \ldots, x_{n,D}\big)^T$.

In this instance, the number of input features $D = 1$.

Rather than working with this fully generalized, multivariate case, let us start by considering a simple linear regression model:

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} = w_0 + w_1 x_1$$

[2] http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

`regression.py` contains the skeleton code for the class `PolynomialRegression`. Objects of this class can be instantiated as `model = PolynomialRegression (m)` where m is the degree of the polynomial feature vector where the feature vector for instance n, $\left(1, x_{n,1}, x_{n,1}^2, \ldots, x_{n,1}^m\right)^T$. Setting $m = 1$ instantiates an object where the feature vector for instance n, $\left(1, x_{n,1}\right)^T$.

(b) **(0 pts)** Note that to take into account the intercept term (w_0), we can add an additional "feature" to each instance and set it to one, e.g. $x_{i,0} = 1$. This is equivalent to adding an additional first column to $\boldsymbol{X}$ and setting it to all ones.

Modify `PolynomialRegression.generate_polynomial_features(...)` to create the matrix $\boldsymbol{X}$ for a simple linear model.

(c) **(0 pts)** Before tackling the harder problem of training the regression model, complete `PolynomialRegression.predict(...)` to predict $\boldsymbol{y}$ from $\boldsymbol{X}$ and $\boldsymbol{w}$.

(d) **(10 pts)** One way to solve linear regression is through gradient descent (GD).

Recall that the parameters of our model are the w_j values. These are the values we will adjust to minimize $J(\boldsymbol{w})$. In gradient descent, each iteration performs the update

$$w_j \leftarrow w_j - 2\alpha \sum_{n=1}^{N} \left(h_{\boldsymbol{w}}(\boldsymbol{x}_n) - y_n\right) x_{n,j} \quad \text{(simultaneously update } w_j \text{ for all } j\text{)}.$$

With each step of gradient descent, we expect our updated parameters w_j to come closer to the parameters that will achieve the lowest value of $J(\boldsymbol{w})$.

- **(0 pts)** As we perform gradient descent, it is helpful to monitor the convergence by computing the cost, *i.e.*, the value of the objective function J. Complete `PolynomialRegression.cost(...)` to calculate $J(\boldsymbol{w})$.

  If you have implemented everything correctly, then the following code snippet should return 40.234.

```
train_data = load_data('regression_train.csv')
model = PolynomialRegression()
model.coef_ = np.zeros(2)
model.cost(train_data.X, train_data.y)
```

- **(0 pts)** Next, implement the gradient descent step in `PolynomialRegression.fit_GD(...)`. The loop structure has been written for you, and you only need to supply the updates to $\boldsymbol{w}$ and the new predictions $\hat{y} = h_{\boldsymbol{w}}(\boldsymbol{x})$ within each iteration.

  We will use the following specifications for the gradient descent algorithm:

  - We run the algorithm for $10,000$ iterations.
  - We terminate the algorithm ealier if the value of the objective function is unchanged across consecutive iterations.
  - We will use a fixed step size.

- **(10 pts)** So far, you have used a default learning rate (or step size) of $\eta = 0.01$. Try different $\eta = 10^{-4}, 10^{-3}, 10^{-2}, 0.0407$, and make a table of the coefficients, number of iterations until convergence (this number will be $10,000$ if the algorithm did not converge in a smaller number of iterations) and the final value of the objective function. How do the coefficients compare? How quickly does each algorithm converge?

(e) **(5 pts)** In class, we learned that the closed-form solution to linear regression is

$$\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

Using this formula, you will get an exact solution in one calculation: there is no "loop until convergence" like in gradient descent.

- **(0 pts)** Implement the closed-form solution `PolynomialRegression.fit(...)`.
- **(5 pts)** What is the closed-form solution? How do the coefficients and the cost compare to those obtained by GD? How quickly does the algorithm run compared to GD?

(f) **(5 pts)** Finally, set a learning rate η for GD that is a function of k (the number of iterations) (use $\eta_k = \frac{1}{1+k}$) and converges to the same solution yielded by the closed-form optimization (minus possible rounding errors). Update `PolynomialRegression.fit_GD(...)` with your proposed learning rate. How long does it take the algorithm to converge with your proposed learning rate?

Polynomial Regression[15 pts]

Now let us consider the more complicated case of polynomial regression, where our hypothesis is

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^T\phi(\boldsymbol{x}) = w_0 + w_1 x + w_2 x^2 + \ldots + w^m x^m.$$

(g) **(0 pts)** Recall that polynomial regression can be considered as an extension of linear regression in which we replace our input matrix $\boldsymbol{X}$ with

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi(x_1)^T \\ \phi(x_2)^T \\ \vdots \\ \phi(x_N)^T \end{pmatrix},$$

where $\phi(x)$ is a function such that $\phi_j(x) = x^j$ for $j = 0, \ldots, m$.

Update `PolynomialRegression.generate_polynomial_features(...)` to create an $m+1$ dimensional feature vector for each instance.

(h) **(5 pts)** Given N training instances, it is always possible to obtain a "perfect fit" (a fit in which all the data points are exactly predicted) by setting the degree of the regression to $N-1$. Of course, we would expect such a fit to generalize poorly. In the remainder of this problem, you will investigate the problem of overfitting as a function of the degree of

the polynomial, m. To measure overfitting, we will use the Root-Mean-Square (RMS) error, defined as

$$E_{RMS} = \sqrt{J(\boldsymbol{w})/N},$$

where N is the number of instances.[3]

Why do you think we might prefer RMSE as a metric over $J(\boldsymbol{w})$?

Implement `PolynomialRegression.rms_error(...)`.

(i) **(10 pts)** For $m = 0, \ldots, 10$, use the closed-form solver to determine the best-fit polynomial regression model on the training data, and with this model, calculate the RMSE on both the training data and the test data. Generate a plot depicting how RMSE varies with model complexity (polynomial degree) – you should generate a single plot with both training and test error, and include this plot in your writeup. Which degree polynomial would you say best fits the data? Was there evidence of under/overfitting the data? Use your plot to justify your answer.

Regularization[5 pts]

Finally, we will explore the role of regularization. For this problem, we will use L_2-regularization so that our regularized objective function is

$$J(\boldsymbol{\theta}) = \frac{1}{2}\sum_{n=1}^{N}(h_{\boldsymbol{\theta}}(\boldsymbol{x}_n) - y_n)^2 + \frac{\lambda}{2}||\boldsymbol{\theta}_{[1:m]}||^2,$$

again optimizing for the parameters $\boldsymbol{\theta}$.

(j) **(0 pts)** Modify `PolynomialRegression.fit(...)` to incorporate L_2-regularization.

(k) **(5 pts)** Use your updated solver to find the coefficients that minimize the error for a tenth-degree polynomial ($m = 10$) given regularization factor $\lambda = 0, 10^{-8}, 10^{-7}, \ldots, 10^{-1}, 10^{0}$. Now use these coefficients to calculate the RMS error (unregularized) on both the training data and test data as a function of λ. Generate a plot depicting how RMS error varies with λ (for your x-axis, let $x = [1, 2, \ldots, 10]$ correspond to $\lambda = [0, 10^{-8}, 10^{-7}, \ldots, 10^{0}]$ so that λ is on a logistic scale, with regularization increasing as x increases). Which λ value appears to work best?

[3] Note that the RMSE as defined is a biased estimator. To obtain an unbiased estimator, we would have to divide by $n - k$, where k is the number of parameters fitted (including the constant), so here, $k = m + 1$.