CSCI 567, Spring'18
Michael Shindler

Problem Set #1

Due: 11:59 pm, February 11, 2018

Instructions

Submission: Assignment submission will be via courses.uscden.net. By the submission date, there will be a folder set up in which you can submit your files. Please be sure to follow all directions outlined here.

You can submit multiple files, but only the last one submitted counts. That means if you finish some problems and want to submit something now, and then a later file when you finish, that's fine. If I were taking the class, that's what I'd do: that way, if I forget to finish the homework or something happens (remember Murphy's Law), I still get credit for what I finished and turned in. Remember, there are no grace days on problem sets, just on programming assignments!

Problem sets must be typewritten or neatly handwritten when submitted; if the grader cannot read your handwriting on a problem, they may elect to grade it as a zero. If it is handwritten, your submission must still be submitted as a PDF.

It is strongly recommended that you typeset with LaTeX and use that to generate a PDF file.

- The file should be named as `firstname_lastname_USCID.pdf` (e.g., Jenny_Tutone_8675309.pdf).
- Do not have any spaces in your file name when uploading it.
- Please include your name and USCID in the header of the report as well.

There are many free integrated LaTeX editors that are convenient to use. Choose the one(s) you like the most. This http://www.andy-roberts.net/writing/latex seems like a good tutorial.

Collaboration: You may discuss with your classmates. However, you need to write your own solutions and submit separately. Also in your report, you need to list with whom you have discussed for each problem. Please consult the syllabus for what is and is not acceptable collaboration. Review the rules on academic conduct in the syllabus: a single instance of plagiarism can adversely affect you significantly more than you could stand to gain.

Notes on notation:

- Unless stated otherwise, scalars are denoted by small letter in normal font, vectors are denoted by small letters in bold font and matrices are denoted by capital letters in bold font.
- The bias term is subsumed in the input vector, so the input vector is actually $x = [x', 1]^T$, unless mentioned otherwise.
- $\|.\|$ means L2-norm unless specified otherwise i.e. $\|.\| = \|.\|_2$

Problem 1 Regularized Linear Regression as Maximum Aposterior (MAP) Estimate (10 points)

In class, we studied regularization as a way to avoid overfitting. Regularization can be interpreted as assuming a prior distribution on the weights $\mathbf{w}$ of regression. In the linear regression model, we are given N independently sampled points

$$\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots (\mathbf{x}_N y_N)\} \tag{1}$$

$\mathbf{x}$ is d-dimensional vector in $\mathbb{R}^d$ and distributed as $p(\mathbf{x})$ and $y \in \mathbb{R}$ is generated by

$$y = \mathbf{w}^T\mathbf{x} + \epsilon, \tag{2}$$

where ϵ is the gaussian noise with mean 0 and variance σ, and $\mathbf{w}$ is an unknown d-dimensional vector (i.e. model parameters) in $\mathbb{R}^d$. We want to estimate the model parameters $\mathbf{w}$ given the data $\mathcal{D}$. In Bayesian perspective, the estimate of $\mathbf{w}$ will be a probability distribution $p(\mathbf{w}|\mathcal{D})$ and so we assume a prior distribution over $\mathbf{w}$ as $\mathcal{N}(0, \alpha\mathbf{I})$ i.e. $p(\mathbf{w}) = \mathcal{N}(0, \alpha\mathbf{I})$.

1.1 Write the expression for $p(y|\mathbf{x}, \mathbf{w})$ in terms of $\mathbf{x}, \mathbf{w}, \sigma$, and $\mathcal{N}$. **(2 points)**

$p(y|\mathbf{x}, \mathbf{w}) = \mathcal{N}(\mathbf{w}^T\mathbf{x}, \sigma)$ (Simple shifting of distribution)
Note: Award points for using either σ or σ^2 as variance

1.2 Write the expression for $p(\mathbf{w}|\mathcal{D})$ in terms of the distributions computed or given so far (*Hint: Use Bayes Rule*) **(2 points)**

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathbf{w})p(\mathcal{D}|\mathbf{w})}{p(\mathcal{D})} = \frac{p(\mathbf{w})\prod_{i=1}^N p((\mathbf{x} = \mathbf{x}_i, y = y_i)|\mathbf{w})}{p(\mathcal{D})} = \frac{p(\mathbf{w})\prod_{i=1}^N p(y = y_i|\mathbf{x} = \mathbf{x}_i, \mathbf{w})p(\mathbf{x} = \mathbf{x}_i)}{p(\mathcal{D})}$$

Note: Students may simplify $p(\mathcal{D})$ further but it is not required

1.3 Show that the MAP estimate $\hat{\mathbf{w}}$ leads to the same objective function as regularized linear regression problem.

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \; p(\mathbf{w}|\mathcal{D}) \tag{3}$$

Hint:

$$\arg\max_{\mathbf{w}} \; p(\mathbf{w}|\mathcal{D}) = \arg\min_{\mathbf{w}} \; -\log p(\mathbf{w}|\mathcal{D})$$

(4 points)

$$\arg\max_{\mathbf{w}} \; p(\mathbf{w}|\mathcal{D}) = \arg\min_{\mathbf{w}} \; -\log p(\mathbf{w}|\mathcal{D})$$

$$= \arg\min_{\mathbf{w}} \; -\log p(\mathbf{w}) - \sum_{i=1}^N \log p(y = y_i|\mathbf{x} = \mathbf{x}_i, \mathbf{w}) - \sum_{i=1}^N \log p(\mathbf{x} = \mathbf{x}_i) + \log p(\mathcal{D})$$

$$= \arg\min_{w} \frac{\mathbf{w}^T\mathbf{w}}{2\alpha} + \sum_{i=1}^N \frac{(y_i - \mathbf{w}^T\mathbf{x}_i)^2}{2\sigma} + C$$

($\log p(\mathbf{x} = \mathbf{x}_i)$ and $\log p(\mathcal{D})$ does not depend on $\mathbf{w}$ and hence is a constant from optimization point of view. So that and other distribution related constants are subsumed in C)

$$\lambda = \sigma/\alpha$$

and deduce that MAP estimate leads to regularized objective

1.4 What happens when we do not assume prior on $\mathbf{w}$ and instead maximize $p(\mathcal{D}|\mathbf{w})$ i.e. find the parameters $\hat{\mathbf{w}}$ which maximize the possibility of occurrence of the given sample $\mathcal{D}$? Show that the objective function in this case will be the same as the unregularized linear regression problem.

(Note: $p(\mathcal{D}|\mathbf{w})$ is called likelihood of data and finding parameters such that likelihood is maximized is maximum likelihood estimation) **(2 points)**

$$\arg\max_{\mathbf{w}} \; p(\mathcal{D}|\mathbf{w}) = \arg\min_{\mathbf{w}} \; -\sum_{i=1}^{N} \log p(y = y_i|\mathbf{x} = \mathbf{x}_i, \mathbf{w}) - \sum_{i=1}^{N} \log p(\mathbf{x} = \mathbf{x}_i)$$

$$= \sum_{i=1}^{N} \frac{(y_i - \mathbf{w}^T\mathbf{x}_i)^2}{2\sigma} + C$$

($\log p(\mathbf{x}_i)$ does not depend on $\mathbf{w}$ and hence is a constant from optimization point of view. So that and other distribution related constants are subsumed in C)

From the above deduce that objective is same as unregularized regression problem.

Problem 2 Convergence of Perceptron Algorithm (10 points)

In class we stated that if the two classes are linearly seperable, Perceptron algorithm will converge.

Recall that in Perceptron algorithm, each example $\mathbf{x}_i$ has a label $y_i \in \{-1, 1\}$ and the algorithm updates weight $\mathbf{w}$ using the following update rule

$$\mathbf{w}_{k+1} = \begin{cases} \mathbf{w}_k + y_k\mathbf{x}_k & \text{if } \; y_k\mathbf{w}_k^T\mathbf{x}_k < 0 \\ \mathbf{w}_k & \text{otherwise} \end{cases} \tag{4}$$

That is after each mistake weights are updated, if the algorithm does not make mistake, weights do not change. Additionally, assume that all the examples are normalized i.e. $\|\mathbf{x}_k\| = \sqrt{\mathbf{x}_k^T\mathbf{x}_k} = 1$. Let $\gamma(\mathbf{w}) = \min_{\mathbf{x}} \frac{|\mathbf{w}^T\mathbf{x}|}{\|\mathbf{w}\|}$. $\gamma(\mathbf{w})$ is also called margin of the classifier $\mathbf{w}$. Convince yourself that $\gamma(\mathbf{w})$ is the smallest distance of any data point from the seperating plane $\mathbf{w}$. Let $\mathbf{w}_{opt}$ be the optimal weights i.e. it linearly separates the data with maximum margin. Note that since data is linearly seperable there will always exist some $\mathbf{w}_{opt}$. Let $\gamma = \gamma(\mathbf{w}_{opt}) = \min_{\mathbf{x}} \frac{|\mathbf{w}_{opt}^T\mathbf{x}|}{\|\mathbf{w}_{opt}\|}$

In this problem, we will show that perceptron algorithm will make finite number of mistakes. The maximum number of mistakes that it can make are γ^{-2}.

2.1 Show that if the algorithm makes a mistake, the update rule moves it in the direction of the optimal weights $\mathbf{w}_{opt}$. Mathematically show that, if $y_k\mathbf{w}_k^T\mathbf{x}_k < 0$

$$\mathbf{w}_{k+1}^T\mathbf{w}_{opt} \geq \mathbf{w}_k^T\mathbf{w}_{opt} + \gamma\|\mathbf{w}_{opt}\| \tag{5}$$

Hint: Consider $(\mathbf{w}_{k+1} - \mathbf{w}_k)^T\mathbf{w}_{opt}$ *and consider the property of* $\mathbf{w}_{opt}$ **(2.5 points)**

$$\mathbf{w}_{k+1} = \mathbf{w}_k + y_k\mathbf{x}_k \tag{6}$$

$$\mathbf{w}_{k+1}^T\mathbf{w}_{opt} = \mathbf{w}_k^T\mathbf{w}_{opt} + y_k\mathbf{x}_k^T\mathbf{w}_{opt} \tag{7}$$

Since, $y_k \mathbf{x}_k \mathbf{w}_{opt}$ is positive (by definition of w_{opt}), so by def of γ

$$\mathbf{w}_{k+1}^T \mathbf{w}_{opt} \geq \mathbf{w}_k^T \mathbf{w}_{opt} + \gamma \|\mathbf{w}_{opt}\| \tag{8}$$

$$\tag{9}$$

Note: In this case, we can not really say that update is in the optimal direction, but magnitude of projection of $\mathbf{w}_{k+1}$ on $\mathbf{w}_{opt}$ has increased which means apart from other things, component of weight in optimal direction increases

2.2 Show that the length of updated weights do not increase by large amount. Mathematically show that, if $y_k \mathbf{w}_k^T \mathbf{x}_k < 0$

$$\|\mathbf{w}_{k+1}\|^2 \leq \|\mathbf{w}_k\|^2 + 1 \tag{10}$$

Hint: Consider $\|\mathbf{w}_{k+1}\|^2$ *and substitute* $\mathbf{w}_{k+1}$ **(2.5 points)**

$$\|\mathbf{w}_{k+1}\|^2 = \mathbf{w}_{k+1}^T \mathbf{w}_{k+1} = (\mathbf{w}_k + y_k \mathbf{x}_k)^T (\mathbf{w}_k + y_k \mathbf{x}_k) \tag{11}$$

$$= \|\mathbf{w}_k\|^2 + 2 y_k \mathbf{w}_k^T \mathbf{x}_k + y_k^2 \mathbf{x}_k^T \mathbf{x}_k \tag{12}$$

Input ($\mathbf{x}_k$) norm is 1 and algorithm has made a mistake so $y_k \mathbf{w}_k^T \mathbf{x}_k \leq 0$

$$\|\mathbf{w}_{k+1}\|^2 = \|\mathbf{w}_k\|^2 + 2 y_k \mathbf{w}_k^T \mathbf{x}_k + y_k^2 \mathbf{x}_k^T \mathbf{x}_k \leq \|\mathbf{w}_k\|^2 + 1 \tag{13}$$

2.3 Assume the initial weights $\mathbf{w}_0$ to be all zeros. Using results from problem 2.1 and 2.2, deduce that after M mistakes

$$M \times \gamma \leq \|\mathbf{w}_{k+1}\| \leq \sqrt{M} \tag{14}$$

Hint: $\mathbf{a}^T \mathbf{b} \leq \|\mathbf{a}\| \|\mathbf{b}\|$ *and use the telescopic sum.* **(2.5 points)**

By repeatedly applying results from problem 2.1 for 1 to M mistakes and summing them up.

$$\mathbf{w}_{k+1}^T \mathbf{w}_{opt} \geq \mathbf{w}_0^T \mathbf{w}_{opt} + M\gamma \|\mathbf{w}_{opt}\|$$

Since $w_0 = 0$

$$\mathbf{w}_{k+1}^T w_{opt} \geq M\gamma \|\mathbf{w}_{opt}\|$$

Due to Cauchy Schwartz inequality,

$$M\gamma \|\mathbf{w}_o pt\| \leq \mathbf{w}_{k+1}^T \mathbf{w}_{opt} \leq \|\mathbf{w}_{k+1}\| \|\mathbf{w}_{opt}\|$$

$$M\gamma \leq \|\mathbf{w}_{k+1}\|$$

Similarly, use results of problem 2.2 repeatedly and sum to them all to conclude

$$\|\mathbf{w}_{k+1}\|^2 \leq \|\mathbf{w}_0\|^2 + M$$

Since, initial weights are 0, conclude that

$$\|\mathbf{w}_{k+1}\|^2 \leq M$$

2.4 Using result of problem 2.3, deduce $M \leq \gamma^{-2}$ **(0.5 points)**

obvious

2.5 What happens if the two classes are not linearly seperable? Explain at what point do the above arguments fail?

(2 points)

- Result of problem 2.1 won't hold, as now we can't argue that $y_k \mathbf{x}_k^T \mathbf{w}$ positive for all $\mathbf{x}_i$. Because the data is not linearly seperable infact there will exist some $\mathbf{x}_i$ for which the term will be negative.
- Also, it could be argued that w_{opt} will not exist as there will be no line seperating the data linearly.

Note: We give points for quoting any of the above

Problem 3 Direction of Linear Discriminant Hyperplane (5 points)

Consider linear discriminant analysis for a two class classification problem given N inputs $\{\mathbf{x}_1 \dots \mathbf{x}_N\}$ and corresponding labels $\{y_1 \dots y_N\}$, $y_i \in \{-1, 1\}$ $\forall i \in \{1 \dots N\}$. We say input $\mathbf{x}_i$ belongs to class $\mathcal{C}_1$ if its label y_i is 1 and it belongs to class $\mathcal{C}_{-1}$ if its label is -1. Mathematically, $\mathcal{C}_1 = \{(\mathbf{x}, y) : y = 1\}$ and $\mathcal{C}_{-1} = \{(\mathbf{x}, y) : y = -1\}$

We try to find a seperating hyperplane $\mathbf{w}^T\mathbf{x}$ such that if input $\mathbf{x}_i$ belongs to $\mathcal{C}_1$ then $\mathbf{w}^T\mathbf{x}_i \geq 0$ and if it belongs to $\mathcal{C}_{-1}$ then $\mathbf{w}^T\mathbf{x}_i \leq 0$. Mathematically, we can define $f(\mathbf{w}) = \sum_{i=1}^{N} y_i \mathbf{w}^T \mathbf{x}_i$ and maximize it under the constraint $\|\mathbf{w}\| = 1$ to find the optimal seperating hyperplane. Note that $f(\mathbf{w})$ can be arbitarily maximized by increasing the magnitude of $\mathbf{w}$. So the constraint $\|\mathbf{w}\| = 1$ (or equivalently, $\|\mathbf{w}\|^2 = 1$) is important.

This can be written as a well defined optimization problem using Lagrange multiplier,

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \sum_{i=1}^{N} y_i \mathbf{w}^T \mathbf{x}_i + \lambda(\mathbf{w}^T\mathbf{w} - 1) \tag{15}$$

Show that

$$\hat{\mathbf{w}} \propto \sum_{i:x_i \in \mathcal{C}_1} \mathbf{x}_i - \sum_{j:x_j \in \mathcal{C}_{-1}} \mathbf{x}_j$$

Note: If you do not know about Lagrange multiplier, consider equation 15 as given and proceed from there.

To find the maximum we find the derivative of $f(w) = \sum_{i=1}^{N} y_i \mathbf{w}^T \mathbf{x}_i + \lambda(\mathbf{w}^T\mathbf{w} - 1)$ and set it to 0.

$$\nabla_w f(\hat{\mathbf{w}}) = \sum_{i=1}^{N} y_i \mathbf{x}_i + \lambda \hat{\mathbf{w}} = 0$$

$$\implies \hat{w} \propto \sum_{i=1}^{N} y_i \mathbf{x}_i = \sum_{i:y_i=1} y_i \mathbf{x}_i + \sum_{i:y_i=-1} y_i \mathbf{x}_i = \sum_{i:x_i \in \mathcal{C}_1} \mathbf{x}_i - \sum_{j:x_j \in \mathcal{C}_{-1}} \mathbf{x}_j$$