CSE517A – Homework 1

M. Neumann

Feb 1 2018

- Please keep your written answers brief and to the point. Incorrect or rambling statements can hurt your score on a question.
- If your hand writing is not readable, we **cannot give you credit**. We recommend you type your solutions in LaTeX and compile a .pdf for each answer. **Start every problem on a new page!**
- This homework is will assess your knowledge of the prerequisites for this course.
- This will be due THU **Feb 15 2018 10am**, with an automatic 3-day extension
- You may work in groups of at most 2
- Submission instructions:
  - Start every problem on a **new page**.
  - Submissions will be exclusively accepted via **Gradescope**. Find instructions on how to get your Gradescope account and submit your work on the course webpage.

Problem 1 (*20 points*) Loss function optimization using gradient descent.

Derive the gradient update (with stepsize c) for your weight vector $\vec{w}$ for each of the following loss functions: (here: $\|\vec{w}\|_2^2 = \vec{w}^\top \vec{w}$ and $|\vec{w}| = \sum_{\alpha=1}^{d} |w_\alpha|$, also λ and C are non-negative constants.)

(a) (*5 pts*) Ridge Regression (*cf. implementation project 1*)

$$\mathcal{L}(\vec{w}) = \sum_{i=1}^{n} (\vec{w}^\top \vec{x}_i - y_i)^2 + \lambda \|\vec{w}\|_2^2$$

(b) (*5 pts*) Lasso Regression:

$$\mathcal{L}(\vec{w}) = \sum_{i=1}^{n} (\vec{w}^\top \vec{x}_i - y_i)^2 + \lambda |\vec{w}|$$

(c) (*5 pts*) Logistic Regression ($y_i \in \{+1, -1\}$, *cf. implementation project 1*)

$$\mathcal{L}(\vec{w}) = \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i \vec{w}^\top \vec{x}_i\right)\right)$$

(d) (*5 pts*) Linear Support Vector Machine ($y_i \in \{+1, -1\}$, *cf. implementation project 1*)

$$\mathcal{L}(\vec{w}) = C \sum_{i=1}^{n} \max\{1 - y_i \vec{w}^\top \vec{x}_i,\ 0\} + \|\vec{w}\|_2^2$$

Problem 2 (*25 points*) Discriminative ML and Newton's Method

Let us re-visit Logistic Regression, however with $y_i \in \{0, 1\}$.

(a) (*5 pts*) Show that with these new labels the objective function can be written as

$$\mathcal{L}(\vec{w}) = -\sum_{i=1}^{n} \Big(y_i \log\left(\text{sigm}(\vec{w}^\top \vec{x}_i)\right) + (1 - y_i) \log\left(1 - \text{sigm}(\vec{w}^\top \vec{x}_i)\right) \Big)$$

(b) (*5 pts*) Show that the gradient of $\mathcal{L}$ can be written as

$$\frac{\partial L}{\partial \vec{w}} = -\sum_{i=1}^{n} \left(y_i - \text{sigm}(\vec{w}^\top \vec{x}_i)\right) \vec{x}_i.$$

(HINT: You can use the fact that $\frac{\partial \text{sigm}(z)}{\partial z} = \text{sigm}(z)(1 - \text{sigm}(z))$.)

(c) (*5 pts*) Let the $n \times n$ diagonal matrix $W_{ii} = \text{sigm}(\vec{w}^\top \vec{x}_i)(1 - \text{sigm}(\vec{w}^\top \vec{x}_i))$ and let $X = [\vec{x}_1, \ldots, \vec{x}_n]^\top$. Show that the Hessian matrix is $H = X^\top W X$. For which examples is W_{ii} large, for which is it small?

(d) (*5 pts*) Prove that the negative log likelihood is convex by showing that Hessian H is positive-semi definite.

(e) (*5 pts*) Write down the update rule for a newton step. Show that if you use the substitution $\vec{z}$ where $z_i = \vec{x}_i^\top \vec{w} + \frac{1}{W_{ii}}(y_i - \text{sigm}(\vec{w}^\top \vec{x}_i))$, you arrive at

$$\vec{w}_{\text{new}} \leftarrow (X^\top W X)^{-1} X^\top W \vec{z}.$$

Problem 3 (*30 points*) Weighted Ridge Regression

Assume that in addition to your data $\{(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)\}$ you also have weights $p_i \geq 0$ for each example. Let your loss function be

$$\mathcal{L}(\vec{w}) = \sum_{i=1}^{n} p_i (\vec{w}^\top \vec{x}_i - y_i)^2 + \lambda \vec{w}^\top \vec{w}$$

(a) (*10 pts*) Rephrase the previous equation in terms of the matrices $X = [\vec{x}_1, \ldots, \vec{x}_n]^\top$, $Y = [y_1, \ldots, y_n]^\top$ and the diagonal matrix $P = \text{diag}([p_1, \ldots, p_n])$ (where the diag operator performs like the Matlab function with the same name.)

(b) (*10 pts*) Derive a closed form solution for $\vec{w}$. (You can use: $\frac{\partial(\vec{w}^\top A)}{\partial \vec{w}} = A$, $\frac{\partial(\vec{w}^\top B \vec{w})}{\partial \vec{w}} = B\vec{w} + B^\top \vec{w}$ and $\vec{w}^\top \vec{w} = \vec{w}^\top I \vec{w}$ where I is the identity matrix.)

(c) (*10 pts*) Let $\lambda = 0$. Look at the result of **Problem 2(e)** Newton's Method. Hold your breath and enjoy the moment of amazement. Why is that algorithm also called *Iteratively Reweighted Least Squares*?