CS 584 - Machine Learning

Instructor: Steve Avsec

Email: savsec@iit.edu

Due: 2/23 at Midnight (Thursday night)

Exercises

0.1 Normal Equations

[5 points] We have used the normal equation a couple of times in lecture.

$$\mathbf{c} = (A^t A)^{-1} A^t \mathbf{y}.$$

We have also used the pseudoinverse

$$\mathbf{c} = A^\dagger \mathbf{y}$$

defined by $A^\dagger = V\Sigma^\dagger U^t$ where $A = U\Sigma V^t$ is the singular value decomposition of A, and $\Sigma^\dagger$ is the diagonal matrix with positive elements (σ_j^{-1}).

Show that $A^\dagger = (A^t A)^{-1} A^t$ if A is full rank.

0.2 Reformulate Ridge and LASSO

[10 points]

1. In lecture, we defined ridge regularization by the minimizer:

   $$L(\mathbf{c}) = \|\mathbf{y} - A\mathbf{c}\|_2^2 + \lambda \|\mathbf{c}\|_2^2.$$

   Show that this is equivalent to the constrained minimization problem:

   $$L(\mathbf{c}) = \|\mathbf{y} - A\mathbf{c}\|_2^2 \tag{1}$$
   $$\text{such that } \|\mathbf{c}\|_2^2 \leq t \tag{2}$$

   Describe how λ and t are related.

2. Similarly, we defined LASSO using the minimizer

   $$L(\mathbf{c}) = \|\mathbf{y} - A\mathbf{c}\|_2^2 + \lambda \|\mathbf{c}\|_1.$$

   Show that this is equivalent to the constrained minimization

   $$L(\mathbf{c}) = \|\mathbf{y} - A\mathbf{c}\|_2^2$$
   $$\text{such that } \|\mathbf{c}\|_1 \leq t$$

   Again describe how λ and t are related.

0.3 Naive Bayes

[10 Points]

1. What key assumption makes the Naive Bayes model naive?
2. Suppose you have a data set with 9500 benign Python scripts and 500 malicious. 200 of the benign scripts import the ast library while 100 of the malicious samples import the ast library. Given that a script imports the ast library, what is the probability that it is malicious?

Implementation

[25 Points] Write a scripts to implement ridge regularization for logistic regression from first principles (you may use NumPy and SciPy, but not library implementations like you would find in Scikit Learn or Statsmodels or any number of R packages) and apply them to the data set given. (The x_j columns are intended to be inputs while the y column is intended to represent the target variable.) Here are some steps that might help.

1. Recall from lecture that

$$P(\mathbf{y}|\mathbf{X}) = \prod_{j=1}^{N} \left(\frac{e^{c_0+\mathbf{c}\cdot\mathbf{x}_j}}{1+e^{c_0+\mathbf{c}\cdot\mathbf{x}_j}} \right)^{n_j} \left(\frac{1}{1+e^{c_0+\mathbf{c}\cdot\mathbf{x}_j}} \right)^{1-n_j}$$

   where $y_j = n_j$ and $\mathbf{x}_j$ denotes the features for that sample. Let $\tilde{\mathbf{c}} = (c_0, \mathbf{c})$. Taking logs of the above formula, we get the log loss function

$$L(\tilde{\mathbf{c}}) = \sum_{j=1}^{N} n_j \, (c_0 + \mathbf{c}\cdot\mathbf{x}_j) - \log(1+e^{c_0+\mathbf{c}\cdot\mathbf{x}_j})$$

2. Add the usual regularization term for the coefficients $\tilde{\mathbf{c}}$ to this log loss.

$$L_r(\tilde{\mathbf{c}}) = L(\tilde{\mathbf{c}}) + \lambda \|\tilde{\mathbf{c}}\|_2^2$$

3. Compute the gradient of this regularized log loss.
4. Implement gradient descent for the regularized log loss function.