CS 475 Machine Learning: Homework 2
Supervised Classifiers 1
Due: Wednesday February 19, 2014, 12:00pm
100 Points Total Version 1.0

Yiran Zhang

1 Analytical (50 points)

1) Decision Tree and Logistic Regression (8 points) Consider a binary classification task (label y) with four features (x):

x_1	x_2	x_3	x_4	y
0	1	1	-1	1
0	1	1	1	0
0	-1	1	1	1
0	-1	1	-1	0

(a) Can this function be learned using a decision tree? If so, provide such a tree (describe each node in the tree). If not, prove it.

This cannot be learned by a decision tree. Since a decision tree will look greedily for the attribute that is going to bring the most information gain but none of the four attributes has positive IG.

$$\begin{aligned} IG(y|x_2) &= H(Y) - H(Y|X_2) \\ &= 1 - \sum_{x_2,y} p(x_2,y) \log \frac{p(x_2,y)}{p(x_2)} \\ &= 1 - 4 \cdot 0.25 \cdot \log 0.5 \\ &= 0 \end{aligned}$$

$$\begin{aligned} IG(y|x_4) &= H(Y) - H(Y|X_4) \\ &= 1 - \sum_{x_4,y} p(x_4,y) \log \frac{p(x_4,y)}{p(x_4)} \\ &= 1 - 4 \cdot 0.25 \cdot \log 0.5 \\ &= 0 \end{aligned}$$

Also for x_1 and x_3, they are the same for all y outcome, so they are providing any information.

(b) Can this function be learned using a logistic regression classifier? If yes, give some example parameter weights. If not, why not.

(c) For the models above where you can learn this function, the learned model may over-fit the data. Propose a solution for each model on how to avoid over-fitting.

2) Stochastic Gradient Descent (8 points) In the programming part of this assignment you implemented Gradient Descent. A stochastic variation of that method (Stochastic Gradient Descent) takes an estimate of the gradient based on a single sampled example, and takes a step based on that gradient. This process is repeated many times until convergence. To summarize:

1. Gradient descent: compute the gradient over all the training examples, take a gradient step, repeat until convergence.
2. Stochastic gradient descent: sample a single training example, compute the gradient over that training example, take a gradient step, repeat until convergence.

In the limit, will both of these algorithms converge to the same optimum or different optimum? Answer this question for both convex and non-convex functions. Prove your answers.

3) Regularizer of Regression (10 points) In linear regression we want to minimize the sum of square loss

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} ||y - X\beta||_2^2 \tag{1}$$

To address overfitting, we might plug in a regularizer $||\beta||_q$ as:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} ||y - X\beta||_2^2 + \lambda||\beta||_q \tag{2}$$

where $||\cdot||_q$ is the q-norm and λ is the regularization parameter

(a) What is the effect on β when $q = 0$, $q = 1$ and $q = 2$? Explain why this is the case.

(b) What is the effect of λ in terms of variance/bias trade-off? How do we usually select a proper λ?

(c) Suppose each example has three features and the corresponding parameters are β_0, β_1 and β_2. If we formulate the regularizer as $||\beta_{\{0,1\}}||_2 + |\beta_2|$, where $\beta_{\{0,1\}}$ is a 2 dimensional vector containing the first two elements of β. Describe the effect of this regularizer.

4) Constructing Generalized linear models. (12 points) Generalized linear models (GLMs), especially logistic regression are heavily used by banks, credit card companies and insurance companies. Actually, when you apply for a credit card, banks may put your information into a logistic regression model to decide whether you are eligible.

(a) The GLMs are closely related to the exponential distribution family, which has the probability density/mass function $f(y;\theta)$ in the form

$$f(y;\theta) = h(y)e^{\eta(\theta)\cdot T(y) - A(\theta)}, \tag{3}$$

where h, η, T, A are some known functions.

Consider a classification or regression problem where we would like to predict the value of some random variable y as a function of x. To derive a GLM for this problem, we will make the following three assumptions about the conditional distribution of y given x and about our model:

1. $y|x;\theta \sim ExponentialFamily(\eta)$. I.e., given x and θ, the distribution of y follows some exponential family distribution, with parameter η.
2. Given x, our goal is to predict the expected value of $T(y)$ given x. In most of our examples, we will have $T(y) = y$, so this means we would like the prediction $h(x)$ output by our learned hypothesis h to satisfy $h(x) = E[y|x]$.
3. The natural parameter η and the inputs x are related linearly: $\eta = \theta^T x$

Derive the expression of logistic regression from the Bernoulli distribution:

$$h_\theta(x) = \frac{1}{1 + \exp(-\theta^T x)} \tag{4}$$

by following the above three assumptions.

(b) The GLMs often contain some transformation, which is non-linear such as the log-odds-ratio transformation in the logistic regression. Why do we still call them "linear"?

5) Convex Optimization (12 points) Jenny at Acme Inc. is working hard on her new machine learning algorithm. She starts by writing an objective function that captures her thoughts about the problem. However, after writing the program that optimizes the objective and getting poor results, she returns to the objective function in frustration. Turning to her colleague Matilda, who took CS 475 at Johns Hopkins, she asks for advice. "Have you checked that your function is convex?" asks Matilda. "How?" asks Jenny.

(a) Jenny's function can be written as $f(g(x))$, where $f(x)$ and $g(x)$ are convex. Prove that $f(g(x))$ is a convex function. (Hint: You may find it helpful to use the definition of convexity. Do not use gradient or Hessian, since f and g may not have them.)

Proof:

$$\begin{aligned} f(g(\lambda x_1 + (1-\lambda)x_2)) &\leq f(\lambda g(x_1) + (1-\lambda)g(x_2)) \\ &\leq \lambda f(g(x_1)) + (1-\lambda)f(g(x_2)) \end{aligned}$$

The first equality is because monotonicity and the convexity of g. The second inequality is because of the convexity.

(b) Jenny realizes that she made an error and that her function is instead $f(x) - g(x)$, where $f(x)$ and $g(x)$ are convex functions. Her objective may or may not be convex. Give examples of functions $f(x)$ and $g(x)$ whose difference is convex, and functions $\bar{f}(x)$ and $\bar{g}(x)$ whose difference is non-convex.

$$f(x) = x, g(x) = 0; \bar{f}(x) = 0\bar{g}(x) = x^2$$

"I now know that my function is non-convex," Jenny says, "but why does that matter?"

(c) Why was Jenny getting poor results with a non-convex function?

Using gradient decent on non-convex function can result in convergence to a local optimum. When the function is not that smooth and has a lot of small hills and bowls everywhere, it's very likely that we are going to reach a local minimum, which is much worse than the global one.

(d) One approach for convex optimization is to iteratively compute a descent direction and take a step along that direction to have a new value of the parameters. The choice of a proper stepsize is not so trivial. In gradient descent algorithm, the stepsize is chosen such that it is proportional to the magnitude of the gradient at the current point. What might be the problem if we fix the stepsize to a constant regardless of the current gradient? Discuss when stepsize is too small or too large.

It might happen that we are bouncing around the optimal point but not reach it. Think about the objective is $f(x) = |x|$ and we are at $x = 0.01$ and the step size is 0.02. When we take the gradient descent, then the current point will be bouncing between -0.01 and 0.01.

2 What to Submit

In each assignment you will submit two things.

1. **Code:** Your code as a zip file named `library.zip`. **You must submit source code (.java files)**. We will run your code using the exact command lines described above, so make sure it works ahead of time. Remember to submit all of the source code, including what we have provided to you.

2. **Writeup:** Your writeup as a **PDF file** (compiled from latex) containing answers to the analytical questions asked in the assignment. Make sure to include your name in the writeup PDF and use the provided latex template for your answers.

Make sure you name each of the files exactly as specified (library.zip and writeup.pdf).

To submit your assignment, visit the "Homework" section of the website (http://www.cs475.org/.)

3 Questions?

Remember to submit questions about the assignment to the appropriate group on the class discussion board: http://bb.cs475.org.