University of Washington
Department of Computer Science and Engineering / Department of Statistics

CSE 599 / Stat 592 Machine Learning (Statistics) for Big Data

Homework 1
Winter 2013

Issued: Tuesday, January 15, 2013 **Due:** Tuesday, January 29, 2013

Suggested Reading: Assigned Readings in Case Study I (see website).
Instructions: The homework consists of two parts: (i) Problems 1.1 to 1.3 cover theoretical and analytical questions and (ii) Problem 1.4 covers data analysis questions. Please submit each portion as *separate* sets of pages with your name and userid (UW student number) on each set. For Part II which involves coding, please print out your code and graphs and attach them to the written part of your homework. Refer to the course webpage for policies regarding collaboration and extensions.
Checkpoint: Problem 1.4 "Warm up" is due in class on Tuesday, January 22 as a checkpoint.

Problem 1.1

(Source: KM Exercise 8.6) **Elementary properties of l_2 regularized logistic regression**
Consider minimizing

$$J(\mathbf{w}) = -l(\mathbf{w}, \mathcal{D}_{\text{train}}) + \lambda \|\mathbf{w}\|_2^2$$

where

$$l(\mathbf{w}, \mathcal{D}) = \sum_j \ln \mathbf{P}(y^j | \mathbf{x}^j, \mathbf{w})$$

is the log-likelihood on data set $\mathcal{D}$, for $y^j \in \{-1, +1\}$. Determine whether the following statements are true or false. Briefly explain.

(a) With $\lambda > 0$ and the features x_k^j linearly separable, $J(\mathbf{w})$ has multiple locally optimal solutions.

(b) Let $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} J(\mathbf{w})$ be a global optimum. $\hat{\mathbf{w}}$ is typically sparse (has many zero entries).

(c) If the training data is linearly separable, then some weights w_j might become infinite if $\lambda = 0$.

(d) $l(\hat{\mathbf{w}}, \mathcal{D}_{\text{train}})$ always increases as we increase λ.

(e) $l(\hat{\mathbf{w}}, \mathcal{D}_{\text{test}})$ always increases as we increase λ.

Problem 1.2

(Source: KM Exercise 8.7) **Regularizing separate terms in 2d logistic regression**

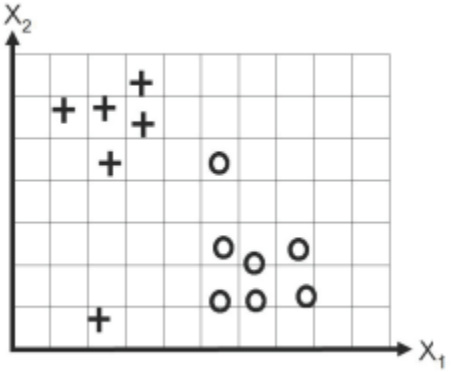

Figure 1 Data for logistic regression question

(a) Consider the data in Figure 1, where we fit the model

$$p(y = 1|\mathbf{x}, \mathbf{w}) = \frac{exp(w_0 + w_1 x_1 + w_2 x_2)}{1 + exp(w_0 + w_1 x_1 + w_2 x_2)}.$$

Suppose we fit the model by maximum likelihood, i.e., we minimize

$$J(\mathbf{w}) = -l(\mathbf{w}, \mathcal{D})$$

where $l(\mathbf{w}, \mathcal{D})$ is the log likelihood on the data set. Sketch a possible decision boundary corresponding to $\hat{\mathbf{w}}$. Is your answer (decision boundary) unique? How many classification errors does your method make on the training set?

(b) Now suppose we regularize only on the w_0 parameter, i.e., we minimize

$$J_0(\mathbf{w}) = -l(\mathbf{w}, \mathcal{D}) + \lambda w_0^2$$

Suppose λ is a very large number, so we regularize w_0 all the way to 0, but all other parameters are unregularized. Sketch a possible decision boundry. How many classification errors does your method make on the data set? Hint: consider the behavior of simple linear regression, $w_0 + w_1 x_1 + w_2 x_2$ when $x_1 = x_2 = 0$.

(c) Now suppose we heavily regularize only the w_1 parameter, i.e.,we minimize

$$J_1(\mathbf{w}) = -l(\mathbf{w}, \mathcal{D}) + \lambda w_1^2$$

Sketch a possible decision boundry. How many classification errors does your method make on the training set?

(d) Now suppose we heavily regularize only the w_2 parameter. Sketch a possible decision boundry. How many classification errors does your method make on the data set?

Problem 1.3

The Count-Min sketch of Cormode and Muthukrishnan is biased. That is, the estimated count $\hat{a}_i$ for element $i \in \{1, \ldots, N\}$ is always higher than (or equal to) the true count a_i. Reminder: The count a_i is the number of times we see element i in the sequence. In this question, you will develop a simple unbiased sketch, *Simple-Count*, (with weaker convergence rates than the Count-Min sketch).

First, we will start with the simplest version of Simple-Count: Let g be a hash function chosen from a family G of independent hashes, such that g maps each i to either $+1$ or -1 with equal probability: [1]:

$$P(g(i) = +1) = P(g(i) = -1) = 1/2.$$

We now define h, the accumulator of our sketch. When we observe element i in the sequence, we simply update:

$$h = h + g(i).$$

Now, if we would like to predict the count for element i, we simply return:

$$\hat{a}_i = h\ g(i).$$

Given this sketch, please answer the following questions:

(a) Let a_i be the true counts for each element i. Express h in terms of the a_i and $g(i)$ only.

(b) What is the expected value of $g(i)$, denoted by $E[g(i)]$?

(c) Prove that $\hat{a}_i = h\ g(i)$ is an unbiased estimate of a_i, i.e., $E[\hat{a}_i] = a_i$. Hint: use linearity of expectations, $E[u+v] = E[u] + E[v]$, and the fact that $g(i)$ and $g(j)$ are independent.

(d) Prove that the variance of our estimate $Var(\hat{a}_i)$ is given by:

$$Var(\hat{a}_i) = \sum_{j \in \{1,\ldots,N\}: j \neq i} a_j^2.$$

Hint: recall that $Var(X) = E[X^2] - (E[X])^2$.

(e) We will now bound the probability of getting a bad estimate. In particular, after n steps, we will say our estimate $\hat{a}_i$ is ϵ-bad if, for $\epsilon > 0$:

$$|\hat{a}_i - a_i| \geq \epsilon n.$$

[1] The randomness arises from the fact that the hash function g is drawn randomly from the family G. Given a hash function g, the mapping $g : \{1, \ldots, N\}$ is deterministic. All expectations, etc. are taken with respect to the distribution of g.

To prove our bound, we will use Chebyshev's inequality: If X is a random variable, and $\alpha > 0$, then:

$$P(|X - E[X]| \geq \alpha) \leq \frac{Var(X)}{\alpha^2}.$$

Use Chebyshev's inequality to prove that the probability δ of getting a bad estimate for $\hat{a}_i$ is bounded by:

$$\delta \leq \frac{Var(\hat{a}_i)}{\epsilon^2 n^2} \leq \frac{1}{\epsilon^2}.$$

(f) The bound in the previous question is going to be vacuous for sufficiently small ϵ. To address this issue, we will expand the number of hash functions in our sketch. Let's introduce a set of k independent hash functions g_j with the same properties as g above. Now, we will create h_j, in analogy to the h function above, for each g_j. When we see element i in the sequence, we will update each h_j by:

$$h_j = h_j + g_j(i).$$

Now, if we would like to predict the count for element i, we simply return the average:

$$\hat{a}_i = \frac{1}{k} \sum_{j=1}^{k} h_j \; g_j(i).$$

For this sketch, prove that:

i. The variance of $\hat{a}_i$ is now bounded by:

$$Var(\hat{a}_i) \leq \frac{n^2}{k}.$$

Hint: The estimates obtained by each hash function are independent.

ii. Use this result and the Chebyshev's inequality as above to prove that for any $\epsilon > 0, \delta > 0$, the probability of getting an ϵ-bad estimate of $\hat{a}_i$ will be lower than δ if we use $k \geq \frac{1}{\delta\epsilon^2}$ hash functions.

Problem 1.4

Logistic Regression for Ads Click Prediction

In this problem, you will train a logistic regression model to predict the Click Through Rate (CTR) on a dataset with ∼1 million examples. The CTR provides a measure of the popularity of an advertisement, and the features we will use for prediction include

attributes of the ad and the user. You will also implement the hashing kernel, where the features are hashed into a smaller space. At the end, there is an extra credit component for implementing multitask logistic regression for personalized CTR prediction (see the "Weinberger, Kilian, et al." paper from the reading list).

Dataset

The dataset we will consider comes from the 2012 KDD Cup Track 2. Here, a user types a query and a set of ads are displayed and we observe which ad was clicked. For example:

1. Alice went to the famous search engine Elgoog, and typed the query "big data".
2. Besides the search result, Elgoog displayed 3 ads each with some short text including its title, description, etc.
3. Alice then clicked on the first advertisement.

This completes a **SESSION**. At the end of this session Elgoog logged 3 records:

Clicked = 1	Depth = 3	Position = 1	Alice	Text of Ad1
Clicked = 0	Depth = 3	Position = 2	Alice	Text of Ad2
Clicked = 0	Depth = 3	Position = 3	Alice	Text of Ad3

In addition, the log contains information about Alice's age and gender. Here is the format of a complete row of our training data:

Clicked	Depth	Position	Userid	Gender	Age	Text Tokens of Ad

Let's go through each field in detail:

- "Clicked" is either 0 or 1, indicating whether the ad is clicked.
- "Depth" takes a value in $\{1, 2, \ldots, \}$ specifying the number of ads displayed in the session.
- "Position" takes a value in $\{1, 2, \ldots, Depth\}$ specifying the rank of the ad among all the ads displayed in the session.
- "Userid" is an integer id of the user.
- "Age" takes a value in $\{1, 2, 3, 4, 5, 6\}$, indicating different ranges of a user's age: '1' for (0, 12], '2' for (12, 18], '3' for (18, 24], '4' for (24, 30], '5' for (30, 40], and '6' for greater than 40.
- "Gender" takes a value in $\{-1, 1\}$, where -1 stands for male and 1 stands for female.

- "Text Tokens" is a comma separated list of token ids. For example: "15,251,599" means "token_15", "token_251", and "token_599". (Note that due to privacy issues, the mapping from token ids to words is not revealed to us in this dataset, e.g., "token_32" to "big".)

Here is an example that illustrates the concept of features "Depth" and "Position". Suppose the list below was returned by Elgoog as a response to Alice's query. The list has *depth* = 3. "Big Data" has *position* = 1, "Machine Learning" has *position* = 2 and so forth.

Big Data
Machine Learning
Cloud Computing

Here is a sample from the training data:

0 | 2 | 2 | 280151 | 1 | 2 | 0,1,154,173,183,188,214,234,26,3,32,36,37,4503,51,679,7740,8,94

The test data are in the same format except that they do not have the first label field, which is stored in a separate file named "test_label.txt". Some data points do not have user information. In these cases, the userid, age, and gender are set to zero.

Feature Representation

In class, we simply denote

$$x^t = [x_1^t, \ldots, x_d^t] \tag{1}$$

as an abstract feature vector. In the real world, however, constructing the feature vector requires some thought.

- First of all, not everything in the data should be treated as a feature. In this dataset, "Userid" should not be treated as feature.
- Similarly, we cannot directly use the list of token ids as features in Eq. 1 since the numbers are arbitrarily assigned and thus meaningless for the purposes of regression. Instead, we should think of the list of token ids $L \equiv [l_1, l_2, l_3, \ldots]$ as a compact representation of a sparse binary vector $\mathbf{b}$ where $\mathbf{b}[i] = 1 \quad \forall i \in L$. It is important to think in terms of the binary representation but implement the code using a compact representation.
- As for the rest of the features: "Depth", "Position", "Age", and "Gender", they are scalar variables, so please use their original value as the feature.

Accessing and Processing the Data

(a) Download "clickprediction_data.zip" from the course website.

(b) After unzipping the folder, there should be three files: train.txt, test.txt and test_label.txt.

(c) For instructions on setting up Java/Eclipse and using the starter code, please read section at the end of the file.

1. Warm up

We begin by simply assessing various attributes of the dataset, primarily to ensure that it is correctly accessed and parsed.

If you are using the starter code, please complete the functions in "analysis/BasicAnalysis.class". In the starter code, you will find "analysis/DummyLoader.class" as sample code for initializing the dataset, iterating over each row, parsing the text, and printing out results.

(a) Report the average CTR for the training data (Number of clicks / Number of examples).

(b) How many unique tokens are there in the training data? What about the test data? How many tokens appear in both datasets?

(c) How many unique users are there in the training data? What about the test data? How many users appear in both datasets?

2. Stochastic Gradient Descent

Recall that stochastic gradient descent (SGD) performs a gradient descent using a noisy estimate of the full gradient based on just the current example.

(a) Write down the equation for the weight update step. That is, how to update weights w^t using the data point (x^t, y^t), where $x^t \equiv [x_1^t, x_2^t, \ldots, x_d^t]$ is the feature vector for example t, and $y^t \in \{0, 1\}$ is the label.

(b) For stepsizes $\eta = \{0.001, 0.01, 0.05\}$ and without regularization, implement SGD and train the weights by making one pass over the dataset. **Use only one pass over the data on all subsequent questions as well.** For each step size:

- Plot the average loss $\overline{L}$ as a function of the number of steps T, where

$$\overline{L}(T) = \frac{1}{T} \sum_{t=1}^{T} (\hat{y}^t - y^t)^2$$

where $\hat{y^t}$ is the predicted label of example x^t using the weights w^{t-1}. Record the the average loss every 100 steps, e.g. $[100, 200, 300, \dots]$.

- Report the l_2 norm of the weights at the end of the pass.
- Use the weights to predict the CTRs for the test data. Recall that "test_label.txt" contains the labels for the test data. Report the RMSE (root mean square error) of your predicted CTR. Also report the RMSE of the baseline prediction you got from the Warm Up. (Do not expect a huge improvement since the label distribution is biased. Elgoog still makes a huge profit even with a 0.1% improvement in accuracy.)
  Hint: you can use the given Util/EvalUtil.class to compute RMSE.

(c) For $\eta = 0.01$, report the weights for the following features: intercept, "Position", "Depth", "Gender", and "Age". Provide an interpretation of the effect of each feature on the probability of a click based on these inferred weights.

Hint
Java users: You need to complete the "LogisticRegression.class". Ignore the lambda and "performDelayedRegularization()" for now, which will be useful in the next question "Regularization".

Big data is often sparse. In this problem, the feature space is huge (the order is on the size of the entire token vocabulary). Fortunately, you do not need to update every feature for every data point. Why? Because a data point only has a few tokens, and the gradient of w_i will be non-zero if and only if feature i is non-zero. In other words, you just need to update the weights corresponding to the tokens that appear in the current example. Other weights will stay the same. Taking advantage of the data sparsity is one of the key weapons for attacking big data problems.

3. Regularization

Notice that the l_2 norm of the weights in the previous part is not small and keeps growing as we get more and more data. It is necessary to add l_2 regularization to each update step. However, the regularization is not a sparse update. At every step, the regularization affects the weights for all of the features, not just the ones that appeared in the current example. To deal with this issue, we will try to be as lazy as possible. What if at each iteration we just regularize the weights that affect the current example and hope for the best? Unfortunately, this is too lazy because it will be unfair to features that appear frequently.

The trick is to delay the regularization for w_i until we encounter a data point that affects it. Suppose feature i appeared for the first time at time t_1. No regularization of w_i is needed because its value is 0. Then at time t_2, feature i shows up again. You

know that the regularization for w_i was delayed for $t_2 - t_1 - 1$ steps, so its time to let it pay. How much? Each step of the regularization downweights w_i by a factor of $(1 - \eta * \lambda)$, so the total is $(1 - \eta * \lambda)^{t_2 - t_1 - 1}$. To implement the lazy regularization, you need to keep track of the update timestamp for the weights on sparse tokens.

(a) Implement the regularization, and train the weights again using stepsize $\eta = 0.05$ with λ ranging from $0-0.014$ spaced by 0.002, e.g. $[0, 0.002, 0.004, 0.006, \ldots, 0.014]$.

(b) Predict the CTR for the test data and evaluate the RMSE. Plot the RMSE as a function as λ.

Hint
Java users: You need to complete the function "performDelayedRegularization()" in "LogisticRegression.class". "Weights.accessTime" is a map for keeping track of the access time of token weights. For example, "w.accessTime.get(256)" should return the most recent time when the weight for token_256 was updated, or *null* if it's never been updated before.

4. Hashing Kernel

The "Weinberger, Kilian, et al." paper introduces an unbiased hash kernel $\phi : \mathcal{X} \to \mathcal{F}$. The original feature space $\mathcal{X}$ is transformed into a space $\mathcal{F}$ with lower dimension through two hash functions: $h : \mathcal{I} \to \{0, \ldots, m-1\}$, and $\xi : \mathcal{I} \to \{+1, -1\}$, where $\mathcal{I}$ indexes the original feature space $\mathcal{X}$. In this problem, we only ask you to hash the text features, keeping the rest of the features as before. Therefore, $\mathcal{I}$ will be the space of all token ids.

The new feature vector (for the text features) $\phi(x)$ will be an m-dimensional array, where the $\phi(x)_i = \sum_{j:h(j)=i} \xi(j) X_j$. Now, we can run the same SGD algorithm in the hashed feature space. The sparse updating and lazy regularization tricks still apply.

Train the weights in the hashed feature space with $m = \{97, 12289, 1572869\}$, $\lambda = 0.001$ and stepsize $\eta = 0.01$. Report the RMSE of the predicted CTRs for all 3 cases.

Hint
Java users: Complete the "HashDataInstance.class" and "LogisticRegressionWithHashing.class". The starter code has two hash functions in "util/HashUtil.class" where you can use as h and ξ. Ignore the personalized flag. Make sure the runtime does not depend on the size of the hash space m.

5. Extra Credit: Personalization

If you have read and understood the "Weinberger, Kilian, et al." paper in its entirety, you can implement a personalized version of CTR prediction. It's just a few lines of code to change: Instead of hashing each feature once, you hash it again with the userid. The rest remains the same.

Implement the personalized logistic regression with hashing. Train the weights using $\eta = 0.01$, $m = 12289$, and $\lambda = 0.001$.

(a) Report the RMSE on the test data (including all users).

(b) Report the RMSE just based on the subset of users who appear both in the test and training data.

Instructions for starter code and setup

(a) Eclipse is a good editor for programming Java. Download Eclipse Classic 4.2.1 at:

http://www.eclipse.org/downloads/

To install, just unzip the downloaded file. Double click the eclipse executable to launch.

(b) To import the starter code, go to the menu File → Import. Under general, select Existing Projects into Workspace, and click Next. Choose "Select archive file", and find "stubs.zip" that you downloaded from the course website. Click Finish.

(c) If you use the starter code, here are files you need to print out and attach to the end of your writeup:

- BasicAnalysis.java
- HashedDataInstance.java
- LogisticRegression.java
- LogisticRegressionWithHashing.java

General advice for Java users

(a) If you get a Nan, either you divided something by zero or the $exp(w^T x)$ overflowed.

(b) Be careful when dividing an integer. Java performs rounding for integer division. For example: x = 1; x/2; gives you zero. Use x/2.0; instead. When you have two integer variables, cast one into double.

(c) Use map.containsKey(key) before asking a value from a map. Or use "Integer" and "Double" object to store the return from map.get(key), and check null. int x = map.getKey(y) will throw an exception if the key y does not exist.

(d) If you are using the starter code, remember to call Dataset.reset() after every pass of the data.

(e) If you get "out of heap space error", you probably need a larger heap space for jvm. In Eclipse, go to the menu run → run configuration. On the left panel, select the application you just ran, on the right panel, select the Arguments next to Main. Type -Xmx1g on the second input box (VM arguments). This will ensure 1G heap space, which should be enough for this homework.