CS 5875 Applied Machine Learning

Homework 1

Weiming Zhang, Yan Jiang

I. Objective

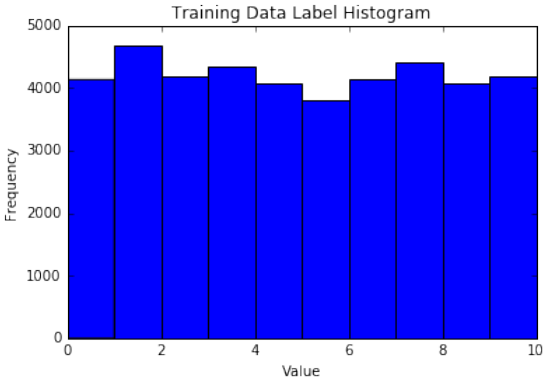
Working on KNN classification method and logistic regression.

II. Problem 1: Digit Recognizer

b) Write a function to display an MNIST digit. Display one of each digit

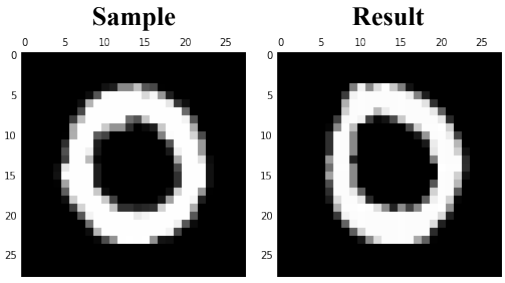
See source code function **displayDigit()**.

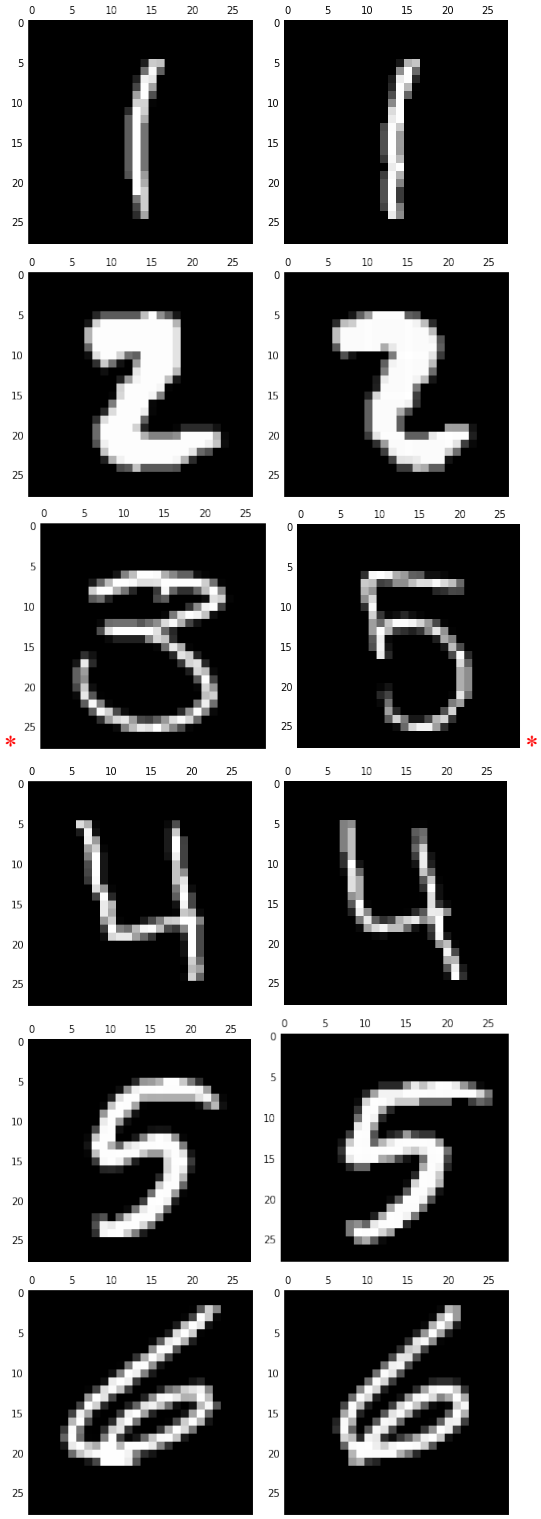
c) Examine the prior probability of the classes in the training data. Is it uniform across the digits? Display a normalized histogram of digit counts. Is it even?

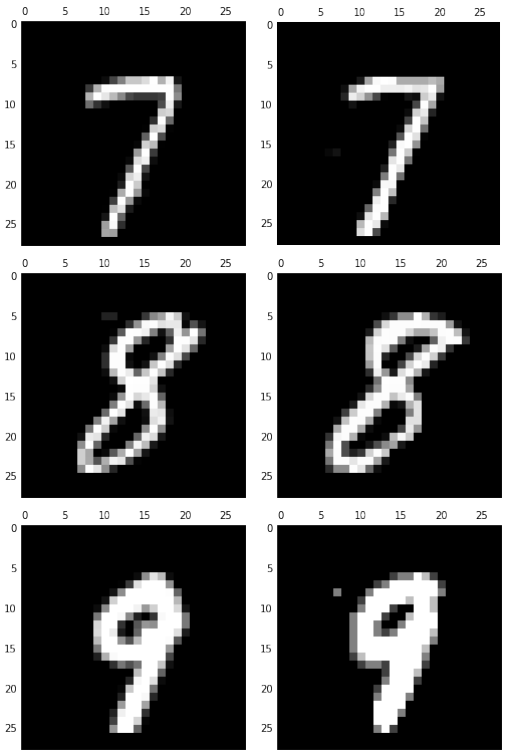


The distribution of the prior probability of the classes in the training data is uniform across the digits. It's evenly distributed. See code for this part in **showLabelDistribution()**.

d) Pick one example of each digit from your training data. Then, for each sample digit, compute and show the best match (nearest neighbor) between your chosen sample and the rest of the training data. Use L2 distance between the two images' pixel values as the metric. This probably won't be perfect, so add an asterisk next to the erroneous examples.

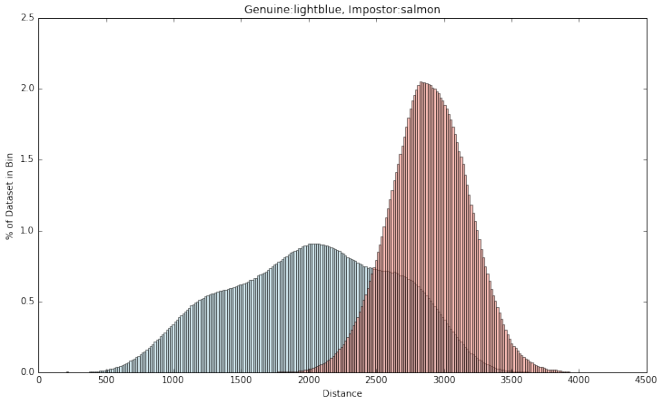






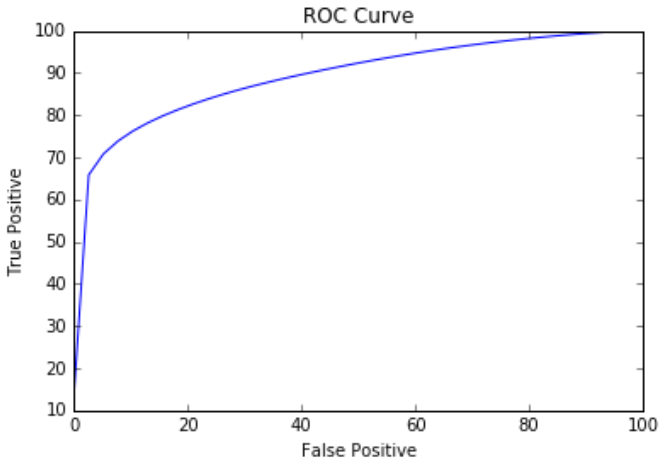
Note for digit 3 the prediction is incorrect (marked with an asterisk). See code for this part in **classify()** in the first panel.

e) *Consider the case of binary comparison between the digits 0 and 1. Ignoring all the other digits, compute the pairwise distances for all genuine matches and all impostor matches, again using the L2 norm. Plot histograms of the genuine and impostor distances on the same set of axes.*



See code for this part in **genuineAndImpostor()**.

f) *Generate an ROC curve from the above sets of distances. What is the equal error rate? What is the error rate of a classifier that simply guesses randomly?*



Equal error rate is around 18%
For random guess equal error rate is 50%
See code for this part in **rocCurve()**.

g) *Implement a K-NN classifier.*

See code for this implementation in the **second panel** of the **Question1.ipynb**

h) *Using the training data for all digits, perform 3 fold cross-validation on your K-NN classifier and report your average accuracy.*

K = 3
Average: 96.55%
See code for this part in **threefold()**.

i) *Generate a confusion matrix (of size 10 × 10) from your results. Which digits are particularly tricky to classify?*

```
[1408    0    0    0    0    1    2    0    0    0]
[   0 1573    2    0    1    0    0    0    0    1]
[  14   10 1291    3    0    2    2   23    6    2]
[   2    6   13 1385    0   13    0    7   10    6]
[   0   13    0    0 1260    0    7    1    1   34]
[   5    2    0   23    1 1208   17    0    5   13]
[   8    2    0    0    1    4 1343    0    0    0]
[   1   12    3    0    3    0    0 1440    0   18]
[   7   25    4   12    5   19    6    4 1272   15]
[   7    3    2   10   10    4    0   24    4 1359]
```

From the matrix we see that digits such as 3, 5 and 9 are relatively tricky to classify.
See code at **confusionMatrix().**

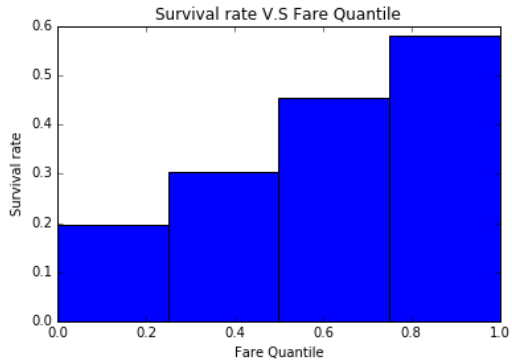
j) *Train your classifier with all of the training data, and test your classifier with the test data. Submit your results to Kaggle.*

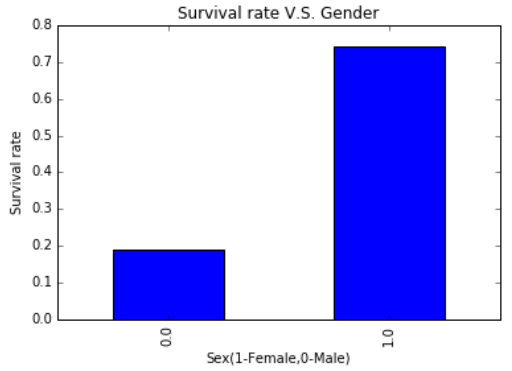
Submitted with accuracy: 96.929%

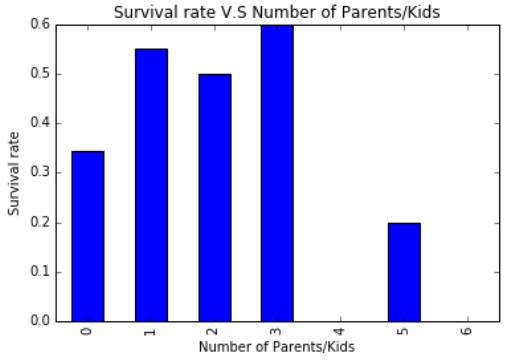
III. Problem 2: The Titanic Disaster

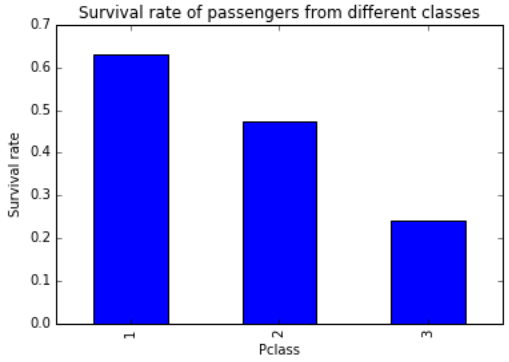
b) *Using logistic regression, try to predict whether a passenger survived the disaster. You can choose the features (or combinations of features) you would like to use or ignore, provided you justify your reasoning.*

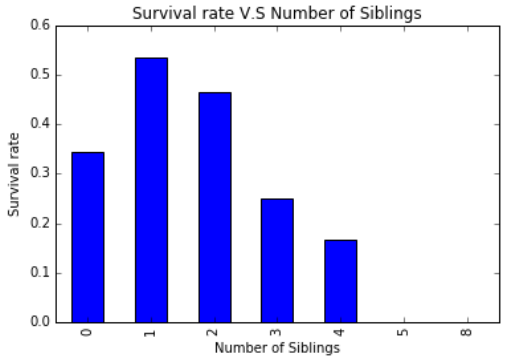
Here are some plots that help us determine the features to choose for classification.











Cross validation accuracy = 80.6%
See code in **Question2.ipynb**

c) *Train your classifier using all of the training data, and test it using the testing data. Submit your results to Kaggle.*

Submitted with accuracy: 75.1%

Applied Machine Learning
Homework 1 - Written Exercises

Yan Jiang, Weiming Zhang

September 14, 2016

Exercise 1. Variance of a sum. Show that the variance of a sum is *var [X +Y]= var [X]+var [Y]+2cov[X,Y]*, where *cov[X,Y]* is the covariance between random variables X and Y .

Proof.

$$
\begin{aligned}
var[X+Y] &= E\left(\left(X+Y-E\left(X+Y\right)^2\right)\right) \\
&= E\left(X+Y\right)^2 - \left(E\left(X+Y\right)\right)^2 \\
&= E\left(X^2+2XY+Y^2\right) - \left(E\left(X+Y\right)\right)^2 \\
var[X] &= E\left(X^2\right) - \left(E\left(X\right)\right)^2 \\
var[Y] &= E\left(Y^2\right) - \left(E\left(Y\right)\right)^2 \\
cov[X+Y] &= E\left(\left(X-E\left(X\right)\right)\left(Y-E\left(Y\right)\right)\right) \\
&= E\left(XY-E\left(X\right)Y-XE\left(Y\right)+E\left(X\right)E\left(Y\right)\right) \\
var[X]+var[Y]+2cov[X+Y] &= E\left(X^2\right)-\left(E\left(X\right)\right)^2+E\left(Y^2\right)-\left(E\left(Y\right)\right)^2 \\
&\quad +2E\left(XY\right)-2E\left(X\right)E\left(Y\right)-2E\left(X\right)E\left(Y\right)+2E\left(X\right)E\left(Y\right) \\
&= E\left(X^2\right)+2E\left(XY\right)+E\left(Y^2\right)-\left(\left(E\left(X\right)\right)^2+\left(E\left(Y\right)\right)^2+2E\left(X\right)E\left(Y\right)\right) \\
&= E\left(X^2+2XY+Y^2\right)-\left(E\left(X\right)+E\left(Y\right)\right)^2 \\
&= E\left(X^2+2XY+Y^2\right)-\left(E\left(X+Y\right)\right)^2 \\
&= var[X+Y]
\end{aligned}
$$

□

Exercise 2. Bayes rule for medical diagnosis (Source: Koller) After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease, and that the test is 99% accurate (i.e., the probability of testing positive given that you have the disease is 0.99, as is the probability of testing negative given that you do not have the disease). The good news is that this is a rare disease, striking only one in 10,000 people. What are the chances that you actually have the disease? (Show your calculations as well as giving the final result.)

Solution:

Let θ indicate the case that a person do have this disease, and let x represent the case that a person is tested positive for this disease.
Here we have:

$$\begin{aligned} P(\theta) &= 0.0001 \\ P(x \mid \theta) &= 0.99 \end{aligned}$$

Thus,

$$\begin{aligned} P(\theta \mid x) &= \frac{P(\theta) P(x \mid \theta)}{P(x)} \\ &= \frac{P(\theta) P(x \mid \theta)}{P(x \mid \theta) P(\theta) + P\left(x \mid \overline{\theta}\right) P\left(\overline{\theta}\right)} \\ &= \frac{0.0001 \times 0.99}{0.0001 \times 0.99 + 0.01 \times 0.9999} \\ &= 0.98 \end{aligned}$$

Exercise 3. Gradient and Hessian of log-likelihood for logistic regression.

(a) Let $\sigma(a) = \frac{1}{1+e^{-}a}$ be the sigmoid function. Show that $\frac{d\sigma(a)}{da} = \sigma(a)(1-\sigma(a))$

Solution:

$$\begin{aligned} \frac{d\sigma(a)}{da} &= (-1) \cdot \frac{1}{(1+e^{-a})^2} \cdot e^{-a} \cdot (-1) \\ &= \frac{e^{-a}}{(1+e^{-a})^2} \\ \sigma(a)(1-\sigma(a)) &= \frac{1}{1+e^{-a}} \left(1 - \frac{1}{1+e^{-a}}\right) \\ &= \frac{1}{1+e^{-a}} \cdot \frac{e^{-a}}{1+e-a} \\ &= \frac{e^{-a}}{(1+e^{-a})^2} \end{aligned}$$

Thus,

$$\frac{d\sigma(a)}{da} = \sigma(a)(1-\sigma(a))$$

(b) Using the previous result and the chain rule of calculus, derive the expression for the gradient of the log likelihood given in HTF Eqn. 4.21.

Solution:

The likelihood of N observation is

$$L(\beta) = \prod_{i}^{N} p\left(x_i;\beta\right)^{y_i}\left(1-p\left(x_i;\beta\right)\right)^{1-y_i}$$

The log likelihood is

$$l(\beta) = \sum_{i=1}^{N}\left[y_i log\left(x_i;\beta\right) + \left(1-y_i\right) log\left(1-p\left(x_i;\beta\right)\right)\right]$$

$$\frac{dl\left(\beta\right)}{d\beta} = \sum_{i=1}^{N}\left[y_i\frac{1}{p\left(x_i;\beta\right)}\cdot\frac{dp\left(x_i;\beta\right)}{d\beta} + \left(1-y_i\right)\cdot\frac{1}{1-p\left(x_i;\beta\right)}\cdot\left(-1\right)\cdot\frac{dp\left(x_i;\beta\right)}{d\beta}\right]$$

Since we have:

$$\frac{dp\left(x_i;\beta\right)}{d\beta} = p\left(x_i;\beta\right)\left(1-p\left(x_i;\beta\right)\right)x_i$$

Thus,

$$\begin{aligned}\frac{dl\left(\beta\right)}{d\beta} &= \sum_{i=1}^{N}\left[y_i\frac{1}{p\left(x_i;\beta\right)}p\left(x_i;\beta\right)\left(1-p\left(x_i;\beta\right)\right)x_i + \left(1-y_i\right)\frac{1}{1-p\left(x_i;\beta\right)}\left(-1\right)p\left(x_i;\beta\right)\left(1-p\left(x_i;\beta\right)\right)x_i\right]\\ &= \sum_{i=1}^{N}\left[y_i\left(1-p\left(x_i;\beta\right)\right)x_i + \left(1-y_i\right)\left(-1\right)p\left(x_i;\beta\right)x_i\right]\\ &= \sum_{i=1}^{N} x_i\left[y_i - p\left(x_i;\beta\right)\right]\end{aligned}$$

(c) As noted in HTF Eqn. 4.25, the Hessian matrix for the log likelihood can be written (up to a sign) as $\mathbf{X^T W X}$. Prove that this matrix is positive definite.

Proof. Since $\mathbf{W}$ is a $\mathbf{N}\times\mathbf{N}$ diagonal matrix of weights with ith diagonal element $p\left(x_i;\beta\right)\left(1-p\left(x_i;\beta\right)\right)$, for any non zero vector

$$\mathbf{X} = \left[x_1, x_2, ...x_i...\right]^T$$

,

$$\mathbf{X^T W X} = \sum_{i=1}^{N} \left[x_i^2 p\left(x_i; \beta\right) \left(1 - p\left(x_i; \beta\right)\right) \right] > 0$$

Thus, $\mathbf{X^T W X}$ is positive definite.

□