In the Name of God

Introduction to Machine Learning (25737-2)

Problem Set 02

Spring Semester 1402-03

Department of Electrical Engineering

Sharif University of Technology

Instructor: Dr. R. Amiri

Due on Farvardin 31, 1403 at 23:59

(∗) starred problems are optional!

1 Easy peasy lemon squeezy

suppose that we have collected the data of a group of students of ML course so that this data includes two features. The first characteristic, x_1, is equal to the total number of hours that the student has practiced ML. The second characteristic, x_2, is the student's total grade point average(GPA) before taking the ML course. Now, we are going to use a logistic regression model to predict the probability that a student will get a score of 20 in this course. After learning based on these data, the obtained coefficients are equal to $\beta_0 = -5$, $\beta_1 = 0.1$ and $\beta_2 = 0.25$.

1. Calculate the probability that a student with 80 hours of study and an GPA of 18 can get a score of 20 in this course.
2. Consider another student who has a GPA of 16 and wants to get a grade of 20 in this course. According to the model that we have taught, how many hours should he practice in order to achieve this grade with a 90% probability?

2 Multi-class Logistic Regression

One way to extend logistic regression to multi-class (say K class labels) setting is to consider (K-1) sets of weight vectors and define

$$P\left(Y=y_k \mid X\right) \propto \exp \left(w_{k0}+\sum_{i=1}^{d} w_{ki} X_i\right) \text{ for } k=1, \ldots, K-1$$

1. What does this model imply for $P\left(Y=y_K \mid X\right)$?
2. What would be the classification rule in this case?
3. Draw a set of training data with three labels and the decision boundary resulting from a multi-class logistic regression. (The boundary does not need to be quantitatively correct but should qualitatively depict how a typical boundary from multi-class logistic regression would look like.)
4. find log-likelihood if N data sample: $X = \{(x_i, y_i) | i \in 1, ..., N\}$ is observed

5. Now suppose that we add a L2 regularizing term to this objective function:

$$f(X) = log(P(Y|X)) - \lambda \sum_{k=1}^{K} ||w_k||2$$

Now calculate the gradient of the function $f(X)$ with respect to w.

6. Write update rule for the data set X according to the answer of the previous part.and simplify as much as possible and explain what happens to w at each iteration

3 Overfitting and Regularized Logistic Regression

1. Plot the sigmoid function $1/\left(1+e^{-wX}\right)$ vs. $X \in \mathbb{R}$ for increasing weight $w \in \{1, 5, 100\}$. A qualitative sketch is enough. Use these plots to argue why a solution with large weights can cause logistic regression to overfit.

2. To prevent overfitting, we want the weights to be small. To achieve this, instead of maximum conditional likelihood estimation M(C)LE for logistic regression:

$$\max_{w_0,\ldots,w_d} \prod_{i=1}^{n} P\left(Y_i \mid X_i, w_0, \ldots, w_d\right),$$

we can consider maximum conditional a posterior M(C)AP estimation:

$$\max_{w_0,\ldots,w_d} \prod_{i=1}^{n} P\left(Y_i \mid X_i, w_0, \ldots, w_d\right) P\left(w_0, \ldots, w_d\right)$$

where $P\left(w_0, \ldots, w_d\right)$ is a prior on the weights. Assuming a standard Gaussian prior $\mathcal{N}(0, I)$ for the weight vector, derive the gradient ascent update rules for the weights.

4 *Lasso Regression

One of the regularization methods in linear regression problems is the Lasso method. In this method, L1 norm of the model's weights is included in the loss function. This causes the final solution of the problem to become more sparse. In this problem, we will see how the L1 norm term results in more sparsity.

$\mathbf{X} \in \mathbb{R}^{n\times d}$ is a matrix where each row is an observation with d features and we have a total of n observations. $\mathbf{y} \in \mathbb{R}^n$ is our label vector. Assume that $\mathbf{w} \in \mathbb{R}^d$ is the weight vector of our regression model and w^* is the optimum weight vector. Also assume that our data has been whitened, that is: $\mathbf{X}^\top\mathbf{X} = \mathbf{I}$.

In Lasso regression the optimum weight vector is obtained as such:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} J_\lambda(\mathbf{w}),$$

where:

$$J_\lambda = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1.$$

1. First we show that whitening the dataset causes the features to be independent such that w_i^* can be concluded only from the ith feature. To prove this, first show that J_λ can be written as:

$$J_\lambda(w) = g(y) + \sum_{i=1}^{d} f(X_{:,i}, \mathbf{y}, w_i, \lambda),$$

where $X_{:,i}$ is the ith column of $\mathbf{X}$.

2. If $w_i \geq 0$, find w_i.
3. If $w_i < 0$, find w_i.
4. Based on previous sections, on what conditions w_i would equal to zero? How can this conditions be applied?
5. As we know, in Ridge regression, regularization term in the loss function appears as $\frac{1}{2}\lambda\|w\|_2^2$. In this case, when does w_i equal to zero? What is the difference between this case and the previous case?

5 Naive Bayes

Consider a Naive Bayes classification problem with three classes and two features. One of these features comes from a Bernoulli distribution and the other comes from a Gaussian distribution. Features are denoted by random vector $\mathbf{X} = [X_1, X_2]^\top$ and class is denoted by Y.

Prior distribution is:

$$\mathbb{P}[Y = 0] = 0.5,\ \mathbb{P}[Y = 1] = 0.25,\ \mathbb{P}[Y = 2] = 0.25$$

Features distribution is:

$$p_{X_1|Y}(x_1|y = c) = Ber(x_1; \theta_c),$$
$$p_{X_2|Y}(x_2|y = c) = \mathcal{N}(x_2; \mu_c, \sigma_c^2).$$

Also assume that:

$$\theta_c = \begin{cases} 0.5 & \text{if } c = 0 \\ 0.5 & \text{if } c = 1 \\ 0.5 & \text{if } c = 2 \end{cases}, \quad \mu_c = \begin{cases} -1 & \text{if } c = 0 \\ 0 & \text{if } c = 1 \\ 1 & \text{if } c = 2 \end{cases}, \quad \sigma_c^2 = \begin{cases} 1 & \text{if } c = 0 \\ 1 & \text{if } c = 1 \\ 1 & \text{if } c = 2 \end{cases}.$$

1. Find $p_{Y|X_1,X_2}(y|x_1 = 0, x_2 = 0)$ (The answer must be a vector in $\mathbb{R}^3$ where the sum of it's elements equal to 1).
2. Find $p_{Y|X_1}(y|x_1 = 0)$.
3. Find $p_{Y|X_2}(y|x_2 = 0)$.
4. Justify the pattern that you see in your answers.

6 Multivariate Least Squares

So far in class, we have only considered cases where our target variable Y is a scalar value. Suppose that instead of trying to predict a single output, we have a training set with multiple outputs for each example:

$$(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}), i = 1, ..., m, \quad \mathbf{x}^{(i)} \in \mathbb{R}^n, \mathbf{y}^{(i)} \in \mathbb{R}^p$$

Thus for each training example, $\mathbf{y}^{(i)}$ is vector-valued, with p entries. We wish to use a linear model to predict the outputs, as in least squares, by specifying the parameter matrix $\mathbf{\Theta}$ in:

$$\mathbf{y} = \mathbf{\Theta}^\top \mathbf{x}$$

where $\mathbf{\Theta} \in \mathbb{R}^{n\times p}$.

1. The cost function for this case is:

$$J(\mathbf{\Theta}) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{p}((\mathbf{\Theta}^\top \mathbf{x}^{(i)})_j - (\mathbf{y}^{(i)})_j)^2.$$

   Write $J(\mathbf{\Theta})$ in matrix-vector notation (i.e., without using any summations).
   [Hint: Start with the $m \times n$ design matrix:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}^{(1)\top} \\ \mathbf{x}^{(2)\top} \\ \vdots \\ \mathbf{x}^{(m)\top} \end{bmatrix},$$

   and the $m \times p$ target matrix:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}^{(1)\top} \\ \mathbf{y}^{(2)\top} \\ \vdots \\ \mathbf{y}^{(m)\top} \end{bmatrix}.$$

   Then work out how to express $J(\mathbf{\Theta})$ in terms of these matrices.]

2. Find the closed form solution for $\mathbf{\Theta}^*$ which minimizes $J(\mathbf{\Theta})$. This is the equivalent to the normal equations for the multivariate case.

3. Suppose instead of considering the multivariate vectors $\mathbf{y}^{(i)}$ all at once, we instead compute each variable $(\mathbf{y}_j^{(i)}$ separately for each $j = 1, ..., p$. In this case, we have p individual linear models, of the form:

$$y_j^{(i)} = \boldsymbol{\theta}_j^\top \mathbf{x}^{(i)}, j = 1, ..., p$$

   (So here, each $\boldsymbol{\theta}_j \in \mathbb{R}^n$). How do the parameters from these p independent least squares problems compare to the multivariate solution?

7 LDA Vs. LR

Consider the one-dimensional feature $\mathbf{X}$ and the two-class response $\mathbf{Y}$. We want to show that classification using linear discriminant analysis is equivalent to using a linear regression model. Specifically, if the sample x^i belongs to the first class we have $\mathbf{Y}^i = -\frac{n}{n_1}$ and if it belongs to the second class we have $\mathbf{Y}^i = \frac{n}{n_2}$. n_1 and n_2 are the number of observations from the first and second classes, and also $n = n_1 + n_2$.

1. Using the definition of the discriminant function, show that LDA labels the sample $\mathbf{X}$ of the second class if:

$$\frac{\hat{\mu_1}-\hat{\mu_2}}{\hat{\sigma^2}}X > \frac{\hat{\mu_2}^2-\hat{\mu_1}^2}{2\hat{\sigma^2}} - log\frac{n_2}{n_1}$$

   And otherwise, it is labeled as first class.

2. Show that the least squares estimate of $\hat{\beta_1}$ in the linear regression model $\mathbf{Y}^i = \hat{\beta_0} + \hat{\beta_1}X^i$ is equal to a multiple (which is only dependent of n) of LDA coefficient for $\mathbf{X}$ in part 1.

3. using the previous results, conclude that LDA is equal to comparing the output of linear model $\hat{\beta_0} + \hat{\beta_1}X$ with a constant.

8 QDA

consider samples in the form of $x_1^i = r\ cos(\theta_i)$ and $x_2^i = r\ sin(\theta_i)$. For the first group $r = 3$ and for the second group $r = 5$. We have 16 samples from each group and $\theta_i = i * \frac{\pi}{8}$ for $1 \leq i \leq 16$. Find the decision boundary using QDA classifier.

9 Hana

Consider a regression problem with all variables and response having mean zero and standard deviation one. Suppose also that each variable has identical absolute correlation with the response:

$$\frac{1}{N} |\langle \mathbf{x}_j, \mathbf{y} \rangle| = \lambda, j = 1, \ldots, p.$$

Let $\hat{\beta}$ be the least-squares coefficient of $\mathbf{y}$ on $\mathbf{X}$, and let $\mathbf{u}(\alpha) = \alpha \mathbf{X} \hat{\beta}$ for $\alpha \in [0, 1]$ be the vector that moves a fraction α toward the least squares fit u. Let RSS be the residual sum-of-squares from the full least squares fit. (a) Show that

$$\frac{1}{N} |\langle \mathbf{x}_j, \mathbf{y} - \mathbf{u}(\alpha) \rangle| = (1 - \alpha)\lambda, j = 1, \ldots, p,$$

and hence the correlations of each $\mathbf{x}_j$ with the residuals remain equal in magnitude as we progress toward $\mathbf{u}$. (b) Show that these correlations are all equal to

$$\lambda(\alpha) = \frac{(1 - \alpha)}{\sqrt{(1 - \alpha)^2 + \frac{\alpha(2-\alpha)}{N} \cdot RSS}} \cdot \lambda,$$

and hence they decrease monotonically to zero.

10 Ridge Regression

1. Show that the ridge regression estimate is the mean (and mode) of the posterior distribution, under a Gaussian prior $\beta \sim N(0, \tau^2 \mathbf{I})$, and Gaussian sampling model $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I})$. Find the relationship between the regularization parameter λ in the ridge formula, and the variances τ^2 and σ^2.

2. Show that the ridge regression estimates can be obtained by ordinary least squares regression on an augmented data set. We augment the centered matrix $\mathbf{X}$ with p additional rows $\sqrt{\lambda}\mathbf{I}$ and augment $\mathbf{y}$ with p zeroes. By introducing artificial data having response value zero, the fitting procedure is forced to shrink the coefficients toward zero.

11 Estimate

Let $x_1, x_2, ...x_N$ be a sequence of independent random variables from normal distribution with variance σ and mean μ.
Answer the following questions (assume you know the amount of variance.)

1. Compute the estimator MLE for the mean μ
2. Compute the estimator MAP for the mean μ. Assume that the prior distribution of the mean is from a normal distribution with mean μ and variance β^2.
3. Check how the results of the first two parts change as the value of N increases towards infinity.