In the Name of God

Introduction to Machine Learning (25737-2)

Problem Set 02

Spring Semester 1401-02

Department of Electrical Engineering

Sharif University of Technology

Instructor: Dr. S. Amini

Due on Ordibehesht 10, 1402 at 23:55

1 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 vetor. 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 prevous 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 differencec between this case and the previous case?

2 Bayesian Analysis of Exponential Distribution

A car's lifetime can be modeled as exponential random variable X with parameter θ such that $p_X(x;\theta) = \theta e^{-\theta x}$ where $\theta > 0, x \geq 0$.

1. show that MLE for θ is $\hat{\theta} = \frac{1}{\bar{X}} = \frac{1}{\frac{1}{N}\sum_{i=1}^{N} X_i}$.
2. Assume we have gathered three data points $X_1 = 5, X_2 = 6, X_3 = 4$. Caculate MLE for θ.
3. Assume that Θ is a random variable and we have a prior knowledge that Θ comes from a distribution, which is $\Theta \sim Exp(\lambda)$. Choose $\hat{\lambda}$ in a way that $\mathbb{E}[\Theta] = \frac{1}{3}$.
4. Find the posterior distribution $p_{\Theta|\mathcal{D}}(\theta|\mathcal{D};\hat{\lambda})$.
5. Find $\mathbb{E}[\Theta|\mathcal{D},\hat{\lambda}]$.

3 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.

4 Desicion Boundary

Assume $p_X(x|Y = j) = \mathcal{N}(x;\mu_j,\sigma_j)$ such that $j = 1, 2$ and $(\mu_1 = 0, \sigma_1^2 = 1), (\mu_2 = 1, \sigma_2^2 = 10^6)$. Also assume that the probability of each class is equal ($\mathbb{P}[Y = 1] = \mathbb{P}[Y = 2] = 0.5$)

1. Find the decision boundary $R_1 = x : p_X(x|Y = 1) \geq p_X(x|Y = 2)$ and draw it.
2. Now let $\sigma_2^2 = 1$. Once again find R_1 and draw it.

5 Newton's Method as Solver For Linear Regression Problem

In this problem, we will prove that if we use Newton's method solve the least squares optimization problem, then we only need one iteration to converge to $\boldsymbol{\theta}^*$.

1. Find the Hessian of the cost function:

$$J(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{\theta}^\top \mathbf{x}^{(i)} - y^{(i)})^2.$$

2. Show that the first iteration of Newton's method gives us $\boldsymbol{\theta}^* = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$, the solution to our least squares problem.

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 $\boldsymbol{\Theta}$ in:

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

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

1. The cost function for this case is:

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

   Write $J(\boldsymbol{\Theta})$ 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 $J(\boldsymbol{\Theta})$ in terms of these matrices.]

2. Find the closed form solution for $\boldsymbol{\Theta}^*$ which minimizes $J(\boldsymbol{\Theta})$. 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 Choosing A Proper Mapping

We have a classification problem where we have a feature vector $\mathbf{x} \in \mathbb{R}^2$ and two classes (binary classification). Our data is represented as matrices below:

$$\mathbf{X} = \begin{bmatrix} 4 & 0 \\ 3 & 1 \\ 2 & 0 \\ 3 & -1 \\ 6 & 0 \\ 3 & 3 \\ 0 & 0 \\ 3 & -3 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ -1 \\ -1 \end{bmatrix}.$$

Find a proper mapping $\Phi(\mathbf{x})$ which maps $\mathbf{x}$ to a higher dimensional space such that our data is linearly seperable in the introduces space.

(Hint: Draw the data points!)