Homework 4 for #70240413

"Statistical Machine Learning"

Instructor: Prof. Jun Zhu

June 10, 2015

1 Sparse Learning

1.1 Problem 1

For a quadratic function $Q(\boldsymbol{w}) = \mathcal{L}(\boldsymbol{w}; \boldsymbol{X})$, please show the following two optimization problems are equivalent for some values of λ and t, where $\Omega(\boldsymbol{w}) = \|w\|_1$.

$$\min_{\boldsymbol{w}} \quad Q(\boldsymbol{w}) + \lambda \cdot \Omega(\boldsymbol{w}) \tag{1}$$

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & Q(\boldsymbol{w}) \\ s.t. \quad & \Omega(\boldsymbol{w}) \leq t \end{aligned} \tag{2}$$

1.2 Problem 2

Implementation of Lasso.

1.2.1 Lasso Model

Consider the problem

$$\min_{\boldsymbol{\omega} \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\boldsymbol{w}\|_2^2 + \lambda \|\boldsymbol{w}\|_1, \quad \mathbf{y} \in \mathbb{R}^n, \ \mathbf{X} \in \mathbb{R}^{n \times p}, \tag{3}$$

where $\mathbf{X}$ is the feature matrix. Each row of $\mathbf{X}$ is a data vector and each column of $\mathbf{X}$ represents a feature. $\mathbf{y}$ represents the vector of all outputs. $\boldsymbol{w}$ is the feature weight.

1.2.2 Solver

Many algorithms can be used to solve the Lasso problem, such as subgradient descent, proximal methods, coordinate descent, etc. In this section, you are required to use proximal method. Please give your derivation

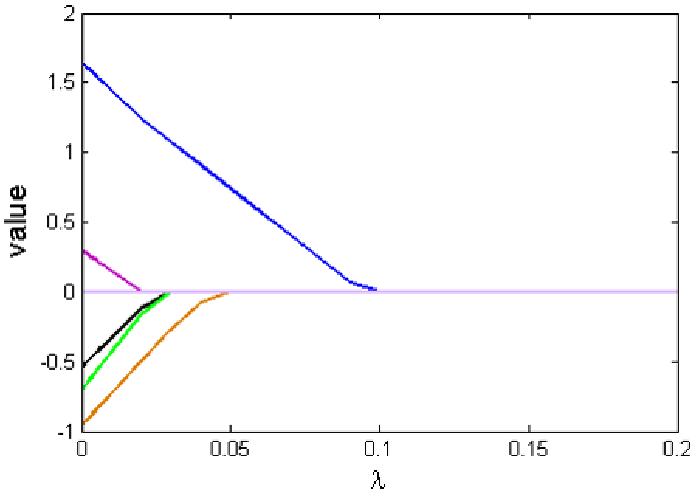

Figure 1: Regularization path of lasso

of the solver and implement it.

1.2.3 Bonus (optional)

You can implement subgradient descent(SGD) or Coordinate Descent (CD) on lasso and give some comparison with the proximal method, in terms of time efficiency and solutions.

1.2.4 Data

The diabetes dataset is provided[1]. In this dataset, the matrices x_train and x_test correspond to $\mathbf{X}$ matrix in Eqn. 3. The matrices y_train and y_test correspond to $\mathbf{y}$ in Eqn. 3.

1.2.5 Evaluation&Tips

Draw regularization path [1] (variation of each dimension in $\boldsymbol{\omega}$ when tuning λ). You may need to tune λ to get the sparsity and please see Fig. 1 as an example.

2 Dirichlet Process

2.1 Problem 1

Consider the Chinese restaurant process CRP(α), compute the expected number of tables occupied when there are n seated customers.

[1] `http://www.cse.msu.edu/~cse847/assignments/diabetes.mat`

2.2 Problem 2

The Pitman-Yor process is closely related to the Dirichlet Process. Recall that in the stick-breaking construction for the Dirichlet Process, we define an infinite sequence of Beta random variables as follows:

$$\beta_i \sim Beta(1, \alpha_0), \quad i = 1, 2, \ldots \tag{4}$$

Then we define an infinite sequence of mixing proportions as follows:

$$\begin{aligned} \pi_1 &= \beta_1 && (5) \\ \pi_k &= \beta_k \prod_{j<k} (1 - \beta_j), \quad k = 2, 3, \ldots && (6) \end{aligned}$$

Under the Pitman-Yor process, the infinite sequence of Beta random variables is defined as

$$\beta_k \sim Beta(1 - d, \alpha + kd), \quad k = 1, 2, \ldots \tag{7}$$

where $0 \leq d < 1$ is a discount parameter. As in the Dirichlet Process, we complete the description of the Pitman-Yor process via

$$\begin{aligned} G &\sim \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} && (8) \\ \theta_i | G &\sim G && (9) \end{aligned}$$

Based on above definitions, please show that the expectation of the total number of occupied tables in the Chinese restaurant scales as $\mathcal{O}(\alpha n^d)$ under the Pitman-Yor process PY(d, α, G_0).

NOTE: This result shows that for natural phenomena that follow power-law distribution, the Pitman-Yor process may be a better choice for a prior than the Dirichlet Process.

2.3 Problem 3

Implementation of Gibbs sampling for Dirichlet process mixture.

2.3.1 DP Mixture Model

You are supposed to use DP Mixture to do clustering on the data $\mathcal{D} = \{x_i\}_{i=1}^N$.

For this dataset, we have $K(\leq N)$ components and membership indicator $\{z_i\}_{i=1}^N$. It should be noted that K might change during the iterations. Data points at each component follows a Normal distribution with parameter $\phi_i = (\mu_i, \Sigma_i), i = 1, 2, ..., K$, $p(x_i | z_i = k, \mu_k, \Sigma_k) \sim \mathcal{N}(\mu_k, \Sigma_k)$. Then the DP mixture model has the following posterior distribution:

$$p(\phi, z | \mathcal{D}) \propto p_0(z) p_0(\phi) p(\mathcal{D} | \phi, z) \tag{10}$$

For the prior of z, the *Chinese Restaurant Process* representation is used. For the prior of ϕ, *Normal-Inverse-Wishart* prior(NIW) is usually chosen as the conjugate prior.

For the Gibbs sampling method, you can either collapse ϕ or not. For simplicity, you can fix $\Sigma = I$ and you only need to set the prior of μ as normal prior $\mathcal{N}(0, \sigma^2 I)$ instead of the NIW prior. For more details, please see [2].

2.3.2 Dataset

You can generate the mixture of Gaussian by yourself, for example, altogether 300 data points and 3 mixtures with equal data size. The gaussian components are taken as 2-dimensional using the parameter $\mu_1 = (2.4, 2), \mu_2 = (-1.8, 2.4), \mu_3 = (-0.2, -2.6)$ and all covariance matrices are set as I.

2.3.3 Evaluation

You need to give some visualization of the clustering. You also need to investigate the following two measures:

- Difference between actual number of clusters K and expected number of clusters $E_{p_0}[K(n, \alpha)]$:

  $D(K; \alpha) = K - \mathrm{E}_{p_0}[K(N)]$ where $\mathrm{E}_{p_0}[K(N, \alpha)]$ is what you get from 2.1

- Mahalanobis distance of all data points to their centers:

  $D_M(\mathcal{D}; \mathbf{z}, \boldsymbol{\phi}) = \sum_{i=1}^{n} |(\mathbf{x}_i - \mu_{z_i})^\top \Sigma^{-1} (\mathbf{x}_i - \mu_{z_i})|^{\frac{1}{2}}$

Please report the curves of the 2 measures during the whole Gibbs sampling process.

References

[1] Tibshirani R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society. Series B (Methodological), 1996: 267-288.

[2] Neal R M. Markov chain sampling methods for Dirichlet process mixture models[J]. Journal of computational and graphical statistics, 2000, 9(2): 249-265.