Homework 1 for #70240413, "Statistical Machine Learning"

Instructor: Prof. Jun Zhu

March 14, 2019

Requirements:

- We recommend that you typeset your homework using appropriate software such as LaTeX. If you submit your handwritten version, please make sure it is cleanly written up and legible. The TAs will not invest undue effort to decrypt bad handwritings.

- We have programming tasks in each homework. Please submit the source code together with your homework. Please include experiment results using figures or tables in your homework, instead of asking TAs to run your code.

- There are optional problems in the assignments. We will give bonus points to those who succeed in solving theses problems.

- Please finish your homework independently.

1 Mathematics Basics

Choose one problem from the 1.1 and 1.2. A bonus would be given if you finished the both.

1.1 Optimization

Use the Lagrange multiplier method to solve the following problem:

$$
\begin{aligned}
\min_{x_1,x_2} \quad & x_1^2 + x_2^2 - 1 \\
s.t. \quad & x_1 + x_2 - 1 = 0 \\
& x_1 - 2x_2 \geq 0
\end{aligned}
\tag{1}
$$

1.2 Calculus

The gamma function is defined by (assuming $x > 0$)

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du. \tag{2}$$

(1) Prove that $\Gamma(x+1) = x\Gamma(x)$.
(2) Also show that

$$\int_0^1 \mu^{a-1} (1-\mu)^{b-1} d\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{3}$$

Note that this result implies that the *Beta distribution* is normalized.

Choose one problem from the following 1.3 and 1.4. A bonus would be given if you finished the both.

1.3 Probability

Suppose $\lambda \sim \Gamma(p|\alpha, \beta)$ and $x|\lambda \sim Poisson(x|\lambda)$. Show that $\lambda|x \sim \Gamma(\lambda|\alpha+x, \beta+1)$, which implies that the Gamma distribution can serve as a conjugate prior to the Poisson distribution.

1.4 Stochastic Process

We toss a fair coin for a number of times and use H(head) and T(tail) to denote the two sides of the coin. Please compute the expected number of tosses we need to observe a first time occurrence of the following consecutive pattern

$$H, \underbrace{T, T, \cdots, T}_{k}.$$

2 SVM

2.1 From Primal to Dual

Consider the regression problem with training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ ($\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$). $\epsilon > 0$ denotes a fixed small value. Derive the dual problem of the following primal problem of linear SVM (Please use α_i, $\hat{\alpha}_i$, β_i, and $\hat{\beta}_i$ as the Lagrange multipliers):

$$
\begin{aligned}
\min_{\boldsymbol{w}, b, \boldsymbol{\xi}, \hat{\boldsymbol{\xi}}} \quad & \frac{1}{2} \|\boldsymbol{w}\|^2 \quad + \quad C \sum_{i=1}^N (\xi_i + \hat{\xi}_i) \\
s.t. \quad y_i \quad & \leq \quad \boldsymbol{w}^\top \boldsymbol{x}_i + b + \epsilon + \xi_i \ , i = 1, \ldots, N \\
y_i \quad & \geq \quad \boldsymbol{w}^\top \boldsymbol{x}_i + b - \epsilon - \hat{\xi}_i \ , i = 1, \ldots, N \\
\xi_i \quad & \geq \quad 0 \quad \forall i = 1, \ldots, N \\
\hat{\xi}_i \quad & \geq \quad 0 \quad \forall i = 1, \ldots, N
\end{aligned}
$$

2.2 Finding Support Vectors (Optional)

State the corresponding set of KKT conditions. Now please argue from the KKT conditions why the following hold:

$$\alpha_i = 0 \Rightarrow \boldsymbol{w}^\top \boldsymbol{x}_i + b + \epsilon - y_i \geq 0$$
$$0 < \alpha_i < C \Rightarrow \boldsymbol{w}^\top \boldsymbol{x}_i + b + \epsilon - y_i = 0$$
$$\alpha_i = C \Rightarrow \boldsymbol{w}^\top \boldsymbol{x}_i + b + \epsilon - y_i \leq 0$$
$$\hat{\alpha}_i = 0 \Rightarrow \boldsymbol{w}^\top \boldsymbol{x}_i + b - \epsilon - y_i \leq 0$$
$$0 < \hat{\alpha}_i < C \Rightarrow \boldsymbol{w}^\top \boldsymbol{x}_i + b - \epsilon - y_i = 0$$
$$\hat{\alpha}_i = C \Rightarrow \boldsymbol{w}^\top \boldsymbol{x}_i + b - \epsilon - y_i \geq 0$$

Please show that either α_i or $\hat{\alpha}_i$ (or both) must be zero. What are the support vectors?

3 IRLS for Logistic Regression

For a binary classification problem $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ ($\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \{0,1\}$), the probabilistic decision rule according to "logistic regression" is

$$P_{\boldsymbol{w}}(y|\boldsymbol{x}) = \frac{\exp(y\boldsymbol{w}^\top \boldsymbol{x})}{1 + \exp(\boldsymbol{w}^\top \boldsymbol{x})}. \tag{4}$$

And hence the log-likelihood is

$$\mathcal{L}(\boldsymbol{w}) = \log \prod_{i=1}^N P_{\boldsymbol{w}}(y_i|\boldsymbol{x}_i) \tag{5}$$

$$= \sum_{i=1}^N \left(y_i \boldsymbol{w}^\top \boldsymbol{x}_i - \log(1 + \exp(\boldsymbol{w}^\top \boldsymbol{x}_i)) \right) \tag{6}$$

Please implement the IRLS algorithm to estimate the parameters of logistic regression

$$\max_{\boldsymbol{w}} \ \mathcal{L}(\boldsymbol{w}) \tag{7}$$

and the L2-norm regularized logistic regression

$$\max_{\boldsymbol{w}} \ -\frac{\lambda}{2}\|\boldsymbol{w}\|_2^2 + \mathcal{L}(\boldsymbol{w}), \tag{8}$$

where λ is the positive regularization constant.

You may refer to the lecture slides for derivation details but you are more encouraged to derive the iterative update equations yourself.

Please compare the results of the two models on the "UCI a9a" dataset[1]. The suggested performance metrics to investigate are e.g. prediction accuracies (both on training and test data), number of IRLS iterations, L2-norm of $\|\boldsymbol{w}\|_2$, etc. You may need to test a range of λ values with e.g. cross validation for the regularized logistic regression.

---

[1] http://ml.cs.tsinghua.edu.cn/~wenbo/data/a9a.zip