Assignment 1 for #70240413
“Statistical Machine Learning”

Instructor: Prof. Jun Zhu

March 12, 2017

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 these problems.

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 Calculus

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

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

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

$$\int_0^1 \mu^{a-1}(1-\mu)^{b-1} d\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{2}$$

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

1.2 Optimization

Use the Lagrange multiplier method to solve the following problem:

$$\begin{aligned} \min_{x_1,x_2} \quad & x_1^2 + x_2^2 - 1 \\ s.t. \quad & x_1 + x_2 - 1 = 0 \\ & 2x_1 - x_2 \geq 0 \end{aligned} \tag{3}$$

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

1.3 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

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

1.4 Probability

Suppose $p \sim \text{Beta}(p|\alpha, \beta)$ and $x|p \sim \text{Bernoulli}(x|p)$. Show that $p|x \sim \text{Beta}(p|\alpha + x, \beta + 1 - x)$, which implies that the Beta distribution can serve as a conjugate prior to the Bernoulli distribution.

2 SVM

2.1 From Primal to Dual

Consider the binary classification problem with training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ ($\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \{0, 1\}$). Derive the dual problem of the following primal problem of linear SVM:

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

(Hint: Please note that we explicitly include the offset b here, which is a little different from the simplified expressions in the slides.)

2.2 Finding Support Vectors (Optional)

As you get the dual problem using KKT conditions. Now please argue from KKT conditions why the following hold:

$$\begin{array}{rcl} \alpha_i = 0 & \Rightarrow & y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 \\ 0 < \alpha_i < C & \Rightarrow & y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) = 1 \\ \alpha_i = C & \Rightarrow & y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \leq 1 \end{array}$$

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

$$\begin{aligned} \mathcal{L}(\boldsymbol{w}) &= \log \prod_{i=1}^N P_{\boldsymbol{w}}(y_i|\boldsymbol{x}_i) && (5) \\ &= \sum_{i=1}^N \left(y_i \boldsymbol{w}^\top \boldsymbol{x}_i - \log(1 + \exp(\boldsymbol{w}^\top \boldsymbol{x}_i))\right) && (6) \end{aligned}$$

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.

Hint: You can use the convergence curves as shown in the lecture slides to show the convergence properties of these two methods.

[1] You can find it in the attachment.