Machine Learning: Homework Assignment 2

E4525 Spring 2019,
IEOR, Columbia University

Due: February 8th, 2019

1 Introduction

The following Facts will be helpful to solve the problems below

1.1 Gamma Function

The Gamma function is the generalization of factorial $n!$ to real valued arguments

$$\Gamma(z) = \int_0^\infty \mathrm{t} t^{z-1} e^{-t} \tag{1}$$

it satisfies

$$\Gamma(n) = (n-1)!. \tag{2}$$

A useful property of $\Gamma(z)$ is

$$\Gamma(z+1) = z\Gamma(z). \tag{3}$$

As a useful special case, $\Gamma(1) = 1$.

1.2 Beta Function

The beta function is defined as

$$B(\alpha, \beta) = \int_0^1 \mathrm{d}\theta\, \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{4}$$

it is related to the Gamma function by equation

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{5}$$

2 Problems

1. **Bernoulli Distribution**

   Assume $S \sim \text{Bernoulli}(\theta)$ for a fixed θ so that

   $$P(S = s; \theta) = C\theta^s(1-\theta)^{1-s} \tag{6}$$

   where C is a normalization constant. Show that

   (a) The normalization constant C is 1.

   (b) $\bar{S} = \mathbb{E}[S] = \theta$

   (c) $\text{Var}(S) = \mathbb{E}[(S - \bar{S})^2] = \theta(1-\theta)$

2. **Bernoulli-Beta Distribution**

   Assume $S \sim \text{Bernoulli}(\theta)$ with a conjugate prior for $\theta \sim \text{Beta}(\alpha, \beta)$ so that

   $$\begin{aligned} P(S = s|\theta) &= \theta^s(1-\theta)^{1-s} \\ P_0(\theta) &= C'\theta^{\alpha-1}(1-\theta)^{\beta-1} \end{aligned} \tag{7}$$

   (a) Compute the normalization constant C' of the beta distribution P_0

   (b) Compute $P(S = s)$ the marginal probability of s.

   (c) Compute $\mathbb{E}(S)$, the expected value of S as a function of α and β.

   (d) Compute $\text{Var}(S)$, the variance of s as a function of α and β.

3. **Maximum Likelihood Inference with the Bernoulli Distribution**

   $S \sim \text{Bernoulli}(\theta)$ for a fixed, but unknown θ. We have performed N independent observations of S, and the results where $\hat{N}_1$ observations with $s = 1$, and $\hat{N}_0 = N - \hat{N}_1$ observations with $s = 0$.

   (a) derive an expresion for θ_{ML} the maximum likelihood estimate of θ given $\hat{N}_1$ and $\hat{N}_0$.

   (b) what is $P(s \mid \hat{N}_1, \hat{N}_0)$ the posterior probability distribution for the next observation s, given the current observations?

   (c) Assume we only performed one experiment $N = 1$, and the observation was $s = 1$, so that $\hat{N}_1 = 1$ and $\hat{N}_0 = 0$, what is maximum likelihood estimated probability for the next observation? Does the result seem sensible to you?

4. **Bayesian Inference with the Bernoulli-Beta Distribution**

   As before $S \sim \text{Bernoulli}(\theta)$ with a conjugate prior $\theta \sim \text{Beta}(\alpha, \beta)$. We have performed N observations of S, and the results where $\hat{N}_1$ observations with $s = 1$, and $\hat{N}_0 = N - \hat{N}_1$ observations with $s = 0$.

(a) write an expression for $P(\theta \mid \hat{N}_1, \hat{N}_0; \alpha, \beta)$ the posterior probability distribution of θ. Express the normalization constant in terms of the beta function.

(b) write an expresion $P(s \mid \hat{N}_1, \hat{N}_0; \alpha, \beta)$ for the probability that the next observation be s given the current observations and the prior.

(c) write an expression for $\mathbb{E}(s \mid \hat{N}_1, \hat{N}_0; \alpha, \beta)$ the expected value of the next observation.

(d) Assume we only performed one experiment $N = 1$, and the observation was $s = 1$, so that $\hat{N}_1 = 1$ and $\hat{N}_0 = 0$, what is maximum likelihood estimated probability for the next observation as a function of α and β? Contrast this the result to the maximum likelihood estimate we obtained on the previous problem.

5. **Maximum Likelihood of Gaussian, Linear Model** Assume the following linear relationship

$$y = \sum_{d=1}^{D} x_d \theta_d + \epsilon \tag{8}$$

where

- $y \in \mathcal{R}$
- $x \in \mathcal{R}^D$
- $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is normally distributed with zero mean and a known σ^2 variance.

We obtain N samples of this process $\{y_i, x_{i,d}\}$ where ϵ_i is independent of ϵ_j if $i \neq j$.

(a) Write the expression the probability density $P(y_i|x_{i,d}; \theta)$ of one observation $\{y_i, x_{i,d}\}$

(b) Write the expression of the join density $P(\{y_i\}|\{x_{i,d}\}; \theta)$ of all the observations $\{y_i, x_{i,d}\}_{i=1}^{N}$

(c) Write the expression of the average log likelihood $\hat{l}(\theta; \{y_i, x_{i,d}\})$ of the observations

(d) Write the expression of the log likelihood loss $E(\theta; \{y_i, x_{i,d}\})$.

(e) Compute expressions for $\frac{\partial E}{\partial \theta_d}$.

(f) Write in matrix form the first order equations that $\Theta = \{\theta_d\}$ and must satisfy for E to be a minimum

(g) Compare to last homework set. Do you see a connection?

6. **Non linear relationship between X and Y**

Assume the following non-linear relationship

$$y = \sum_{d=1}^{D} h_d(x)\theta_d + \epsilon \tag{9}$$

where

- $y \in \mathcal{R}$
- $x \in \mathcal{R}$
- $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is normally distributed.
- $h_d(x)$, for $d = 1, \ldots D$ are a set of arbitrary functions of x

(a) Write the probability density for $P(y_i|x_i;\theta)$

(b) Write the max likelihood loss $E(\theta; \{y_i, x_i\})$ in terms of the matrix $H = H_{i,d} = \{h_d(x_i)\}$, and $Y = \{y_i\}$

(c) Write the matrix equation for the max liikelihood estimate $\Theta = \{\theta_d\}$ in terms H and $Y = \{y_i\}$.

(d) Discuss how what you would need to alter (if anything) in the argument above if $x \in \mathcal{R}^C$ is a C dimensional vector.

(e) What would happen to the max likelihood equations for Θ, if one of the functions $h_d(x)$ can be expressed as a linear combination of the others?

7. **Bayesian Inference: Ridge and Lasso Regressions**

Assume the same model of Exercise 5.

Derive the Bayesian loss function E and its gradient $\frac{\partial E}{\partial \theta_d}$ satisfied by the most likely a-posteriori (MAP) Θ value under the following prior assumptions:

(a) Ridge Regression: $P_0(\Theta) = \mathcal{N}(\Theta; 0, \frac{\sigma^2}{\lambda}\mathbb{1})$

(b) Lasso Regression: $P_0(\Theta) = Ce^{-\lambda \sum_d |\theta_d|}$

(c) Elastic Net: $P_0(\Theta) = Ce^{-\lambda_1 \sum_d |\theta_d|}\mathcal{N}(\Theta; 0, \frac{\sigma^2}{\lambda_2}\mathbb{1})$