

Machine Learning: Homework Assignment 2
E4525 Spring 2018,
IEOR, Columbia University

Due: February 2nd, 2018

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

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

it satisfies

$$\Gamma(n) = (n-1)!. \quad (2)$$

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

$$\Gamma(z+1) = z\Gamma(z). \quad (3)$$

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

1.2 Beta Function

The beta function is defined as

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

it is related to the Gamma function by equation

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

2 Problems

1. Bernoulli Distribution

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

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

where C is a normalization constant. Show that

(a) The normalization constant C is 1.

$P(S)$ must add up to one, but S is Bernoulli so it can only takes values 0 or 1, therefore:

$$1 = P(s = 0; \theta) + P(s = 1; \theta) = C\theta + C(1 - \theta) = C \quad (7)$$

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

By definition

$$\mathbb{E}[S] = \sum_{s=0,1} sP(s; \theta) = 1\theta + 0(1 - \theta) = \theta \quad (8)$$

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

First we compute

$$\mathbb{E}(S^2) = 1^2\theta + 0^2(1 - \theta) = \theta \quad (9)$$

Then, using the identity

$$\text{Var}(S) = \mathbb{E}[(S - \bar{S})^2] = \mathbb{E}[S^2] - (\mathbb{E}[S])^2 \quad (10)$$

we find

$$\text{Var}(S) = \theta - \theta^2 = \theta(1 - \theta) \quad (11)$$

2. Bernoulli-Beta Distribution

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

$$\begin{aligned} P(S = s|\theta) &= \theta^s(1 - \theta)^{1-s} \\ P_0(\theta) &= C'\theta^{\alpha-1}(1 - \theta)^{\beta-1} \end{aligned} \quad (12)$$

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

$$1 = \int d\theta P_0(\theta) = C' \int d\theta \theta^{\alpha-1}(1 - \theta)^{\beta-1} \quad (13)$$

but the integral is just the definition of the **Beta function** therefore

$$C' = \frac{1}{B(\alpha, \beta)} \quad (14)$$

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

$$P(S = 1) = \int d\theta P_0(\theta) P(S = 1 | \theta) = \int d\theta P_0(\theta) \theta \quad (15)$$

and, using the formula for $P_0(\theta)$

$$P(S = 1) = \frac{1}{B(\alpha, \beta)} \int d\theta \theta^\alpha (1 - \theta)^{\beta-1} = \frac{B(\alpha + 1, \beta)}{B(\alpha, \beta)} \quad (16)$$

in terms of gamma functions this is

$$P(S = 1) = \frac{\Gamma(\alpha + 1)\Gamma(\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha + \beta + 1)} \quad (17)$$

finally, using the factorization property of the gamma function $\Gamma(z + 1) = z\Gamma(z)$ we find

$$P(S = 1) = \frac{\alpha}{\alpha + \beta} \quad (18)$$

and

$$P(S = 0) = 1 - P(S = 1) = \frac{\beta}{\alpha + \beta} \quad (19)$$

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

As on the previous problem, $\mathbb{E}(S) = P(S = 1)$ therefore

$$\mathbb{E}(S) = \frac{\alpha}{\alpha + \beta} \quad (20)$$

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

As on the previous problem

$$\text{Var}(S) = \frac{\alpha}{\alpha + \beta} \left(1 - \frac{\alpha}{\alpha + \beta} \right) = \frac{\alpha\beta}{(\alpha + \beta)^2} \quad (21)$$

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 expression for θ_{ML} the maximum likelihood estimate of θ given \hat{N}_1 and \hat{N}_0 .

The probability of obtaining the counts \hat{N}_1 and \hat{N}_0 given θ is

$$P(\hat{N}_1, \hat{N}_0; \theta) = \theta^{\hat{N}_1} (1 - \theta)^{\hat{N}_0} \quad (22)$$

its easier to work with logs so the average log likelihood is

$$\hat{l}(\theta; \hat{N}_1, \hat{N}_0) = \frac{\hat{N}_1}{N} \log \theta + \frac{\hat{N}_0}{N} \log(1 - \theta) \quad (23)$$

The first order equation for the maximum are

$$0 = \frac{\partial \hat{l}}{\partial \theta} = \frac{\hat{N}_1}{N\theta} - \frac{\hat{N}_0}{N(1 - \theta)} \quad (24)$$

or, simplifying a bit

$$0 = \hat{N}_1(1 - \theta) - \hat{N}_0\theta \quad (25)$$

therefore, the max likelihood value of θ is

$$\theta_{\text{ML}} = \frac{\hat{N}_1}{\hat{N}_1 + \hat{N}_0} \quad (26)$$

- (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?

We just plug in θ_{ML} on $P(S; \theta)$ and obtain

$$\begin{aligned} P(S = 1; \hat{N}_1, \hat{N}_0) &= \frac{\hat{N}_1}{\hat{N}_1 + \hat{N}_0} \\ P(S = 0; \hat{N}_1, \hat{N}_0) &= \frac{\hat{N}_0}{\hat{N}_1 + \hat{N}_0} \end{aligned} \quad (27)$$

- (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?

$$P(S = 1) = 1.$$

This seems a rather extreme conclusion based on so little data.

There is no right, or wrong answer, but most human beings will not consider that a reasonable inference, and would, intuitively, still want to assign some probability to the case $s = 0$.

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.

Using Bayes theorem

$$P(\theta | \hat{N}_1, \hat{N}_0; \alpha, \beta) = \frac{P(\theta, \hat{N}_1, \hat{N}_0; \alpha, \beta)}{P(\hat{N}_1, \hat{N}_0; \alpha, \beta)} = \frac{P(\hat{N}_1, \hat{N}_0 | \theta; \alpha, \beta) P_0(\theta; \alpha, \beta)}{\int d\theta P(\hat{N}_1, \hat{N}_0 | \theta; \alpha, \beta) P_0(\theta; \alpha, \beta)} \quad (28)$$

substituting

$$P(\theta | \hat{N}_1, \hat{N}_0; \alpha, \beta) = \frac{\theta^{\alpha + \hat{N}_1 - 1} (1 - \theta)^{\beta + \hat{N}_0 - 1}}{\int d\theta \theta^{\alpha + \hat{N}_1 - 1} (1 - \theta)^{\beta + \hat{N}_0 - 1}} = \frac{\theta^{\alpha + \hat{N}_1 - 1} (1 - \theta)^{\beta + \hat{N}_0 - 1}}{B(\alpha + \hat{N}_1, \beta + \hat{N}_0)} \quad (29)$$

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

As before

$$P(s = 1 | \hat{N}_1, \hat{N}_0; \alpha, \beta) = \int d\theta \theta P(\theta | \hat{N}_1, \hat{N}_0; \alpha, \beta) = \frac{\int d\theta \theta^{\alpha + \hat{N}_1} (1 - \theta)^{\beta + \hat{N}_0 - 1}}{B(\alpha + \hat{N}_1, \beta + \hat{N}_0)} \quad (30)$$

using again the definition of the Beta function it simplifies to

$$P(s = 1 | \hat{N}_1, \hat{N}_0; \alpha, \beta) = \frac{B(\alpha + \hat{N}_1 + 1, \beta + \hat{N}_0)}{B(\alpha + \hat{N}_1, \beta + \hat{N}_0)} \quad (31)$$

finally, using the factorization property of the gamma function $\Gamma(z + 1) = z\Gamma(z)$ we find

$$P(s = 1 | \hat{N}_1, \hat{N}_0; \alpha, \beta) = \frac{\alpha + \hat{N}_1}{\alpha + \hat{N}_1 + \beta + \hat{N}_0} \quad (32)$$

and, again

$$P(s = 0 | \hat{N}_1, \hat{N}_0; \alpha, \beta) = 1 - P(s = 1 | \hat{N}_1, \hat{N}_0; \alpha, \beta) = \frac{\beta + \hat{N}_0}{\alpha + \hat{N}_1 + \beta + \hat{N}_0} \quad (33)$$

- (c) write an expression for $\mathbb{E}(s | \hat{N}_1, \hat{N}_0; \alpha, \beta)$ 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

$$y = \sum_{d=1}^D x_d \theta_d + \epsilon \quad (34)$$

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 $\{y_i, x_{i,d}\}$

$$P(y_i|x_{i,d};\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \sum_d x_{i,d}\theta_d)^2} \quad (35)$$

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

$$P(\{y_i\}|\{x_{i,d}\};\theta) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} e^{-\frac{1}{2\sigma^2} \sum_i (y_i - \sum_d x_{i,d}\theta_d)^2} \quad (36)$$

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

$$\hat{l}(\theta; \{y_i, x_{i,d}\}) = -\frac{1}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2 N} \sum_i \left(y_i - \sum_d x_{i,d}\theta_d \right)^2 \quad (37)$$

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

$$E(\theta; \{y_i, x_{i,d}\}) = -\hat{l}(\theta; \{y_i, x_{i,d}\}) = \frac{1}{2} \log 2\pi\sigma^2 + \frac{1}{2\sigma^2 N} \sum_i \left(y_i - \sum_d x_{i,d}\theta_d \right)^2 \quad (38)$$

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

$$\frac{\partial E}{\partial \theta_d} = \frac{1}{\sigma^2 N} \sum_i \left(\sum_{d'=1}^D x_{i,d'}\theta_{d'} - y_i \right) x_{i,d} \quad (39)$$

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

$$(X^T X)\Theta = X^T Y \quad (40)$$

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

The equations for Θ are the same as we derived last week.

The max likelihood loss E is different from the error we defined last week by a scale $\frac{1}{2\sigma^2}$ and an additive constant $\frac{1}{2} \log 2\pi\sigma^2$ that do not depend on Θ .

6. Non linear relationship between X and Y

Assume the following non-linear relationship

$$y = \sum_{d=1}^D h_d(x)\theta_d + \epsilon \quad (41)$$

where

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

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

$$P(y_i|x_{i,d};\theta,b) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \sum_d h_d(x_i)\theta_d)^2} \quad (42)$$

- (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 $Y = \{y_i\}$

$$E(\Theta) = 2 \log 2\pi\sigma^2 + \frac{1}{\sigma^2 N} (Y - H\Theta)^T (Y - H\Theta) \quad (43)$$

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

$$(H^T H)\Theta = H^T Y \quad (44)$$

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

It is practically the same.

h_d would be C -dimensional functions, but once we have the matrix $H_{i,d}$ computed, the dimensionality of x does not matter.

- (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?

The rank of $H^T H$ would be less than D , and some of the parameters θ would be undetermined (they are redundant).

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

$$E = 2 \log 2\pi\sigma^2 + \frac{1}{2\sigma^2 N} (Y - X\Theta)^T (Y - X\Theta) + \frac{\lambda}{2\sigma^2 N} \Theta^T \Theta \quad (45)$$

$$\frac{\partial E}{\partial \theta_d} = \frac{1}{\sigma^2 N} \left\{ \sum_i x_{i,d} \left(\sum_{d'} x_{i,d'} \theta_{d'} - y_i \right) + \lambda \theta_d \right\} \quad (46)$$

or, in matrix form

$$\frac{\partial E}{\partial \Theta} = \frac{1}{\sigma^2 N} \{ X^T (X\Theta - Y) + \lambda \Theta \} \quad (47)$$

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

$$E = 2 \log 2\pi\sigma^2 + \frac{1}{2\sigma^2 N} \sum_i (y_i - \sum_d x_{i,d} \theta_d)^2 + \frac{\lambda}{N} \sum_d |\theta_d| \quad (48)$$

$$\frac{\partial E}{\partial \theta_d} = \frac{1}{\sigma^2 N} \left\{ \sum_i x_{i,d} \left(\sum_{d'} x_{i,d'} \theta_{d'} - y_i \right) \right\} + \frac{\lambda}{N} \text{sgn } \theta_d \quad (49)$$

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

$$\begin{aligned}
E &= 2 \log 2\pi\sigma^2 + \frac{1}{2\sigma^2 N} \sum_i (y_i - \sum_d x_{i,d}\theta_d)^2 \\
&+ \frac{\lambda_1}{N} \sum_d |\theta_d| \\
&+ \frac{\lambda_2}{\sigma^2 N} \sum_d \theta_d^2
\end{aligned} \tag{50}$$

$$\begin{aligned}
\frac{\partial E}{\partial \theta_d} &= \frac{1}{\sigma^2 N} \left\{ \sum_i x_{i,d} \left(\sum_{d'} x_{i,d'} \theta_{d'} - y_i \right) \right\} \\
&+ \frac{\lambda_1}{N} \operatorname{sgn} \theta_d \\
&+ \frac{\lambda_2}{\sigma^2 N} \theta_d
\end{aligned} \tag{51}$$