Homework 1
MLE, MAP estimates; Linear and Logistic Regression

CMU 10-701: Machine Learning (Spring 2017)

OUT: Jan 31
DUE: Feb 10, 11:59 PM
NAME: Mengwen He
ADREW ID: mengwenh

Part A: Multiple Choice Questions

1. For each case listed below, what type of machine learning problem does it belong to?

   (a) Advertisement selection system, which can predict the probability whether a customer will click on an ad or not based on the search history.
   Answer: B. Supervised learning: Regression
   A task, with ads click statistics and search history as input data, outputs the prediction of continuous probability of clicking an ad.

   (b) U.S post offices use a system to automatically recognize handwriting on the envelope.
   Answer: A. Supervised learning: Classification
   A task, with handwriting samples and their labels as input data, outputs the prediction of discrete numbers/letters of a handwriting on the envelope.

   (c) Reduce dimensionality using principal components analysis (PCA).
   Answer: C. Unsupervised learning
   A task, without training data as input, outputs a description of reduced dimensionality.

   (d) Trading companies try to predict future stock market based on current market conditions.
   Answer:

   - A. Supervised learning: Classification
     A task, with current market conditions as input, outputs the prediction of discrete stock market conditions, say bull or bear market.
   - B. Supervised learning: Regression
     A task, with current market conditions as input, outputs the prediction of continuous stock market conditions, say stock price.

   (e) Repair a digital image that has been partially damaged.
   Answer:

   - A. Supervised learning: Classification
     A task, with digital images database as input, outputs the prediction of discrete pixel value in the damaged zone.
   - B. Supervised learning: Regression
     A task, with digital images database as input, outputs the prediction of continuous parameters of a color distribution model to form a discrete patch to cover the damaged zone.
   - C. Unsupervised learning
     A task, without training data as input, outputs a description of a damaged pixel according to its surrounding pixel values, e.g. interpolation or extrapolation.

   Type of machine learning problem:

A. Supervised learning: Classification

B. Supervised learning: Regression

C. Unsupervised learning

2. For four statements below, which one is wrong?

   A. In maximum a posterior (MAP) estimate, data overwhelms the prior if we have enough data.

   B. There are no parameters in non-parametirc models.

   C. $P(X \cap Y \cap Z) = P(Z|X \cap Y)P(Y|X)P(X)$.

   D. Compared with parametric models, non-parameter models are flexible, since they don't make strong assumptions.

   Answer: B. There are no parameters in non-parametirc models. is wrong.
   Non-parametric model still needs parameters to describe the model, but the number of model's parameters is not fixed and will grow with the data size. The non-parametric only means that there is weak assumption on the model's type defined by a fixed number of parameters.

3. There are about 12% people in U.S. having breast cancer during their lifetime. One patient has a positive result for the medical test. Suppose the sensitivity of this test is 90%, meaning the test will be positive with probability 0.9 if one really has cancer. The false positive is likely to be 2%. Then what is the probability this patient actually having cancer based on Bayes Theorem?
   A. 90% B. 86% C. 12% D. 43%
   Answer: B. 86%

   - $P(C=1) = 0.12$
   - $P(T=1|C=1) = 0.90$
   - $P(T=1|C=0) = 0.02$

$$\begin{aligned} P(C=1|T=1) &= \frac{P(T=1|C=1)P(C=1)}{P(T=1|C=1)P(C=1)+P(T=1|C=0)P(C=0)} \\ &= \frac{0.90\times 0.12}{0.90\times 0.12+0.02\times 0.88} \\ &= 0.86 \end{aligned}$$

4. What is the most suitable error function for gradient **descent** using logistic regression?

   A. The negative log-likelihood function

   B. The number of mistakes

   C. The squared error

   D. The log-likelihood function

   Answer: A. The negative log-likelihood function
   The negative log-likelihood function of a logistic regression is a convex function.

Part B, Problem 1: Bias-Variance Decomposition

Consider a p-dimensional vector $\vec{x} \in \mathbb{R}^p$ drawn from a Gaussian distribution with an identity covariance matrix $\Sigma = I_p$ and an unknown mean $\vec{\mu}$, i.e. $\vec{x} \sim \mathcal{N}(\vec{\mu}, I_p)$. Our goal is to evaluate the effectiveness of an estimator $\hat{\vec{\mu}} = f(\vec{x})$ of the mean from only a single sample (i.e. $n = 1$) by measuring its mean squared error $\mathbb{E}[||\hat{\vec{\mu}} - \vec{\mu}||^2]$, where $||\cdot||^2$ is the squared Euclidean norm and the expectation is taken over the data generating distribution.

Note that for any estimator $\hat{\vec{\theta}}$ of a parameter vector $\vec{\theta}$, its mean squared error can be decomposed as:

$$\mathbb{E}[||\hat{\vec{\theta}} - \vec{\theta}||^2] = ||Bias[\hat{\vec{\theta}}]||^2 + trace(Var[\hat{\vec{\theta}}])$$

where,

$$Bias[\hat{\vec{\theta}}] = \mathbb{E}[\hat{\vec{\theta}}] - \vec{\theta} \; and \; Var[\hat{\vec{\theta}}]_{i,i} = Var[\hat{\theta}_i] = \mathbb{E}[(\hat{\theta}_i - \mathbb{E}[\hat{\theta}_j])^2]$$

1. Derive the maximum likelihood estimator:

$$\hat{\vec{\mu}}_{MLE} = \arg\max_{\vec{\mu}} P(\vec{x}|\vec{\mu})$$

Answer:
$\because$ We estimate the mean $\vec{\mu}$ from only a single sample $\vec{x}_1 \sim \mathcal{N}(\vec{\mu}, I_p)$
$\therefore$ The likelihood function is

$$L(\vec{\mu}) = P(\vec{x}_1|\vec{\mu}) = \frac{1}{(2\pi)^{p/2}|I_p|^{1/2}} \exp(-\frac{1}{2}(\vec{x}_1 - \vec{\mu})^T I_p^{-1}(\vec{x}_1 - \vec{\mu}))$$

$\therefore$

$$\hat{\vec{\mu}}_{MLE} = \arg\max_{\vec{\mu}} L(\vec{\mu}) = \vec{x}_1$$

What is its mean squared error?
Answer:
$\because$ The MSE of $\hat{\vec{\mu}}_{MLE}$ is

$$\begin{aligned}\mathbb{E}[||\hat{\vec{\mu}}_{MLE} - \vec{\mu}||^2] &= ||Bias[\hat{\vec{\mu}}_{MLE}]||^2 + trace(Var[\hat{\vec{\mu}}_{MLE}]) \\ &= ||\mathbb{E}[\hat{\vec{\mu}}_{MLE}] - \vec{\mu}||^2 + \textstyle\sum_{i=1}^p \mathbb{E}[(\hat{\mu}_{MLE_i} - \mathbb{E}[\hat{\mu}_{MLE_i}])^2]\end{aligned}$$

$\because \hat{\vec{\mu}}_{MLE} = \vec{x}_1$
$\therefore \mathbb{E}[\hat{\vec{\mu}}_{MLE}] = \mathbb{E}[\vec{x}_1] = \vec{\mu}$
$\therefore$

$$\mathbb{E}[||\hat{\vec{\mu}}_{MLE} - \vec{\mu}||^2] = \sum_{i=1}^p \mathbb{E}[(\vec{x}_{1_i} - \vec{\mu}_i)^2]$$

$\because \vec{x} \sim \mathcal{N}(\vec{\mu}, I_p)$
$\therefore$

$$MSE = \mathbb{E}[||\hat{\vec{\mu}}_{MLE} - \vec{\mu}||^2] = p$$

2. Derive the ℓ_2-regularized maximum likelihood estimator:

$$\hat{\vec{\mu}}_{RMLE} = \arg\max_{\vec{\mu}} \log P(\vec{x}|\vec{\mu}) - \lambda||\vec{\mu}||^2$$

Answer:
$\because$ We estimate the mean $\vec{\mu}$ from only a single sample $\vec{x}_1 \sim \mathcal{N}(\vec{\mu}, I_p)$
$\therefore$ The ℓ_2-regularized log-likelihood function is

$$L(\vec{\mu}) = C - \frac{1}{2}(\vec{x}_1 - \vec{\mu})^T(\vec{x}_1 - \vec{\mu}) - \lambda\vec{\mu}^T\vec{\mu} = C - \frac{1}{2}\vec{x}_1^T\vec{x}_1 + \vec{x}_1^T\vec{\mu} - (\frac{1}{2} + \lambda)\vec{\mu}^T\vec{\mu}$$

$\therefore$

$$\begin{array}{rcl} \frac{\partial L(\vec{\mu})}{\partial \vec{\mu}}\Big|_{\vec{\mu}_{RMLE}} & = & \vec{0} \\ (\vec{x}_1 - (1+2\lambda)\vec{\mu})|_{\vec{\mu}_{RMLE}} & = & \vec{0} \\ \vec{\mu}_{RMLE} & = & \frac{1}{1+2\lambda}\vec{x}_1 \end{array}$$

What is its mean squared error?
Answer:
From question 1, we know that

$$\mathbb{E}[||\hat{\vec{\mu}}_{RMLE} - \vec{\mu}||^2] = ||\mathbb{E}[\hat{\vec{\mu}}_{RMLE}] - \vec{\mu}||^2 + \sum_{i=1}^{p} \mathbb{E}[(\hat{\mu}_{RMLE_i} - \mathbb{E}[\hat{\mu}_{RMLE_i}])^2]$$

$\because \vec{\mu}_{RMLE} = \frac{1}{1+2\lambda}\vec{x}_1$
$\therefore \mathbb{E}[\hat{\vec{\mu}}_{RMLE}] = \mathbb{E}[\frac{1}{1+2\lambda}\vec{x}_1] = \frac{1}{1+2\lambda}\vec{\mu}$
$\therefore$

$$\mathbb{E}[||\hat{\vec{\mu}}_{RMLE} - \vec{\mu}||^2] = (\frac{2\lambda}{1+2\lambda})^2||\vec{\mu}||^2 + (\frac{1}{1+2\lambda})^2 \sum_{i=1}^{p} \mathbb{E}[(\vec{x}_{1_i} - \vec{\mu}_i)^2]$$

$\because \vec{x} \sim \mathcal{N}(\vec{\mu}, I_p)$
$\therefore$

$$MSE = \mathbb{E}[||\hat{\vec{\mu}}_{RMLE} - \vec{\mu}||^2] = (\frac{2\lambda}{1+2\lambda})^2||\vec{\mu}||^2 + (\frac{1}{1+2\lambda})^2 p$$

3. Consider an estimator of the form $\hat{\vec{\mu}}_{SCALE} = c\vec{x}$ where $c \in \mathbb{R}$ is a constant scaling factor. Find the value c^* that minimizes its mean squared error:

$$c^* = \arg\min_c \mathbb{E}[||c\vec{x} - \vec{\mu}||^2]$$

Answer:
From question 2, if we assume $c = \frac{1}{1+2\lambda}$, we can easily get the objective function for $\hat{\vec{\mu}}_{SCALE} = c\vec{x}$:

$$J_{MSE}(c) = (c-1)^2||\vec{\mu}||^2 + c^2 p$$

$\therefore$

$$\begin{array}{rcl} \frac{dJ_{MSE}(c)}{dc}\Big|_{c^*} & = & 0 \\ ((||\vec{\mu}||^2 + p)c - ||\vec{\mu}||^2)\Big|_{c^*} & = & 0 \\ c^* & = & \frac{||\vec{\mu}||^2}{||\vec{\mu}||^2 + p} \end{array}$$

What is the corresponding minimum mean squared error?
Answer:
If $c^* = \frac{||\vec{\mu}||^2}{||\vec{\mu}||^2+p}$, then

$$MSE^* = J_{MSE}(c^*) = (\frac{p}{||\vec{\mu}||^2 + p})^2||\vec{\mu}||^2 + (\frac{||\vec{\mu}||^2}{||\vec{\mu}||^2 + p})^2 p = \frac{||\vec{\mu}||^2 p}{||\vec{\mu}||^2 + p}$$

4. Consider the James-Stein estimator:

$$\hat{\vec{\mu}}_{JS} = \left(1 - \frac{p-2}{||\vec{x}||^2}\right)\vec{x}$$

Note that $\hat{\vec{\mu}}_{JS}$ can be written as $\vec{x} - g(\vec{x})$ where $g(\vec{x}) = \frac{p-2}{||\vec{x}||^2}\vec{x}$. This allows us to separate the mean squared error into three parts:

$$\begin{array}{rcl} \mathbb{E}[||\hat{\vec{\mu}}_{JS} - \vec{\mu}||^2] & = & \mathbb{E}[||\vec{x} - g(\vec{x}) - \vec{\mu}||^2] \\ & = & \mathbb{E}[\vec{x}^T\vec{x} - 2\vec{x}^T\vec{\mu} + \vec{\mu}^T\vec{\mu} + g(\vec{x}^T g(\vec{x}) - 2\vec{x}^T g(\vec{x}) + 2\vec{\mu}^T g(\vec{x})] \\ & = & \mathbb{E}[||\vec{x} - \vec{\mu}||^2] + \mathbb{E}[||g(\vec{x})||^2] - 2\mathbb{E}[(\vec{x} - \vec{\mu})^T g(\vec{x})] \end{array}$$

Furthermore, from Stein's lemma, we know that:

$$\mathbb{E}[(\vec{x}-\vec{\mu})^T g(\vec{x})] = \mathbb{E}\left[\sum_{j=1}^{p} \frac{\partial}{\partial x_j} g_j(\vec{x})\right]$$

- Find $\mathbb{E}[||\vec{x}-\vec{\mu}||^2]$.
  Answer:

$$\begin{aligned}\mathbb{E}[||\vec{x}-\vec{\mu}||^2] &= \mathbb{E}[\textstyle\sum_{i=1}^{p}(x_i-\mu_i)^2] \\ &= \textstyle\sum_{i=1}^{p}\mathbb{E}[(x_i-\mu_i)^2]\end{aligned}$$

  $\because \vec{x} \sim \mathcal{N}(\vec{\mu}, I_p)$
  $\therefore$

$$\mathbb{E}[||\vec{x}-\vec{\mu}||^2] = p$$

- Find $\mathbb{E}[||g(\vec{x})||^2]$. (Hint: your answer will include $\mathbb{E}[||\vec{x}||^{-2}]$)
  Answer:
  $\because g(\vec{x}) = \frac{p-2}{||\vec{x}||^2}\vec{x}$
  $\therefore$

$$\mathbb{E}[||g(\vec{x})||^2] = (p-2)^2\mathbb{E}[\frac{\vec{x}^T\vec{x}}{(||\vec{x}||^2)^2}] = (p-2)^2\mathbb{E}[||\vec{x}||^{-2}]$$

- Show that:

$$\frac{\partial}{\partial x_j} g_j(\vec{x}) = (p-2)\frac{||\vec{x}||^2 - 2x_j^2}{||\vec{x}||^4}$$

  where x_j is the jth element of x and $g_j(\vec{x})$ is the jth element of $g(\vec{x})$.
  Answer:
  $\because g_j(\vec{x}) = \frac{p-2}{||\vec{x}||^2}x_j$
  $\therefore$

$$\begin{aligned}\tfrac{\partial}{\partial x_j} g_j(\vec{x}) &= \tfrac{\partial}{\partial x_j}(\tfrac{p-2}{||\vec{x}||^2}x_j) \\ &= (p-2)(\tfrac{\partial(\sum_{i=1}^{p}x_i^2)^{-1}}{\partial x_j}x_j + \tfrac{1}{||\vec{x}||^2}) \\ &= (p-2)(\tfrac{-2x_j^2}{(\sum_{j=1}^{p}x_j^2)^{-2}} + \tfrac{1}{||\vec{x}||^2}) \\ &= (p-2)\tfrac{||\vec{x}||^2-2x_j^2}{||\vec{x}||^4}\end{aligned}$$

- What is the resulting mean squred error. (Hint: your answer will include $\mathbb{E}[||\vec{x}||^{-2}]$)
  Answer:
  $\because$
  - $\mathbb{E}[||\vec{x}-\vec{\mu}||^2] = p$
  - $\mathbb{E}[||g(\vec{x})||^2] = (p-2)^2\mathbb{E}[||\vec{x}||^{-2}]$
  - $\frac{\partial}{\partial x_j} g_j(\vec{x}) = (p-2)\frac{||\vec{x}||^2-2x_j^2}{||\vec{x}||^4}$

  $\therefore$

$$\begin{aligned}\mathbb{E}[||\hat{\vec{\mu}}_{JS}-\vec{\mu}||^2] &= \mathbb{E}[||\vec{x}-\vec{\mu}||^2] + \mathbb{E}[||g(\vec{x})||^2] - 2\mathbb{E}\left[\textstyle\sum_{j=1}^{p}\tfrac{\partial}{\partial x_j}g_j(\vec{x})\right] \\ &= p + (p-2)^2\mathbb{E}[||\vec{x}||^{-2}] - 2(p-2)\mathbb{E}[\tfrac{p||\vec{x}||^2 - 2\sum_{j=1}^{p}x_j^2}{||\vec{x}||^4}] \\ &= p + (p-2)^2\mathbb{E}[||\vec{x}||^{-2}] - 2(p-2)^2\mathbb{E}[||\vec{x}||^{-2}] \\ &= p - (p-2)^2\mathbb{E}[||\vec{x}||^{-2}]\end{aligned}$$

5. Qualitatively compare these estimators, noting any similarities between them. How does regularization affect an estimator's bias and variance? Which estimator would you choose to approximate $\vec{\mu}$ from

real data about which you have no prior knowledge? How does the data dimensionality p affect your answer?
Answer:

- Similarities:
  - For $\hat{\vec{\mu}}_{RMLE}$, if $\lambda = 0$, $\hat{\vec{\mu}}_{RMLE} = \hat{\vec{\mu}}_{MLE} = \vec{x}_1$
  - For $\hat{\vec{\mu}}_{SCALE}$, if $c = 1$, $\hat{\vec{\mu}}_{SCALE} = \hat{\vec{\mu}}_{MLE} = \vec{x}_1$
  - For $\hat{\vec{\mu}}_{JS}$, if $p = 2$, $\hat{\vec{\mu}}_{JS} = \hat{\vec{\mu}}_{MLE} = \vec{x}_1$
- The regularization increases the bias, but decreases the variance.
- If I have no prior knowledge, I will choose MLE, because its MSE is not related with the prior knowledge of $\vec{\mu}$ and thus is predictable.
- The increase of dimensionality p will increase the MSE except for the James-Stein estimator.
  $\because \mathbb{E}[||\vec{x}||^{-2}] \geq 0$
  $\therefore MSE(p) = -\mathbb{E}[||\vec{x}||^{-2}]p^2 + (4\mathbb{E}[||\vec{x}||^{-2}] + 1)p - 4\mathbb{E}[||\vec{x}||^{-2}]$ must have maximum value.
  $\therefore$ If the dimensionality p is very large, we can choose James-Stein estimator to constrain its MSE level.

Part B, Problem 2: Linear Regression

Suppose we observe N data pairs $\{(x_i, y_i)\}_{i=1}^N$, where y_i is generated by the following rule:

$$y_i = \vec{x}_i^T \vec{\beta} + \epsilon_i$$

where $\vec{x}_i, \vec{\beta} \in \mathbb{R}^d$, and ϵ_i is an i.i.d random noise drawn from the Gaussian Distribution:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

with a known constant σ. We further denote $\vec{Y} = [y_1, y_2, \ldots, y_N]^T$ and $X = [\vec{x}_1, \vec{x}_2, ..., \vec{x}_N]^T$.

Now, we are interested in estimating $\vec{\beta}$ from the observed data.

1. Derive the likelihood function $\mathcal{L}(\vec{\beta})$

   Answer:

   $\because y_i = \vec{x}_i^T \vec{\beta} + \epsilon_i$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

   $\therefore y_i \sim \mathcal{N}(\vec{x}_i^T \vec{\beta}, \sigma^2)$

   $\therefore$

   $$\begin{aligned} \mathcal{L}(\vec{\beta}) &= \textstyle\prod_{i=1}^N P(y_i|\vec{x}_i, \vec{\beta}) \\ &= \textstyle\prod_{i=1}^N (\frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y_i - \vec{x}_i^T \vec{\beta})^2}{2\sigma^2})) \end{aligned}$$

2. Show that the MLE estimator $\hat{\vec{\beta}}_{MLE}$ of $\vec{\beta}$ is equivalent to the solution of the following linear regression problem:

   $$\min_{\vec{\beta}} \frac{1}{2} ||\vec{Y} - X\vec{\beta}||_2^2 \tag{1}$$

   Answer:

   $\because$ We can derive MLE estimator $\hat{\vec{\beta}}_{MLE}$ via the log likelihood:

   $$\begin{aligned} \hat{\vec{\beta}}_{MLE} &= \arg\max_{\vec{\beta}} \log \mathcal{L}(\vec{\beta}) \\ &= \arg\max_{\vec{\beta}} \textstyle\sum_{i=1}^N -\frac{1}{2\sigma^2}(y_i - \vec{x}^T\vec{\beta})^2 \\ &= \arg\min_{\vec{\beta}} \frac{1}{2} \textstyle\sum_{i=1}^N (y_i - \vec{x}^T\vec{\beta})^2 \\ &= \arg\min_{\vec{\beta}} \frac{1}{2} ||\vec{Y} - X\vec{\beta}||_2^2 \end{aligned}$$

   $\therefore$ The MLE estimator is equivalent to the solution of the following linear regression problem:

   $$J^*(\vec{\beta}) = \min_{\vec{\beta}} \frac{1}{2} ||\vec{Y} - X\vec{\beta}||_2^2$$

3. Now we suppose $\vec{\beta}$ is not a deterministic parameter, but a random variable having a Gaussian prior distribution:

   $$p(\vec{\beta}) \sim \mathcal{N}(\vec{0}, \frac{\sigma^2}{2\lambda} I_d)$$

   where I is a $d \times d$ identity matrix and $\lambda > 0$ is a known parameter. Show that the MAP estimation $\hat{\vec{\beta}}_{MAP}$ of $\vec{\beta}$ is equivalent to the solution of the following ridge regression problem:

   $$\min_{\vec{\beta}} \frac{1}{2} ||\vec{Y} - X\vec{\beta}||_2^2 + \lambda||\vec{\beta}||_2^2 \tag{2}$$

   Answer:

   $\because p(\vec{\beta}) \sim \mathcal{N}(\vec{0}, \frac{\sigma^2}{2\lambda} I_d)$

   $\therefore$

   $$P(\vec{\beta}) = \frac{1}{(2\pi)^{d/2} |\frac{\sigma^2}{2\lambda} I_d|^{1/2}} \exp(-\frac{1}{2}\vec{\beta}^T (\frac{\sigma^2}{2\lambda} I_d)^{-1} \vec{\beta}) = \frac{1}{(\frac{\pi}{\lambda})^{d/2} \sigma^d} \exp(-\frac{\lambda}{\sigma^2} \vec{\beta}^T \vec{\beta})$$

$\because$ the posteriori distribution $P(\vec{\beta}|y_i, \vec{x}_i) \propto P(y_i, |\vec{x}_i, \vec{\beta})P(\vec{\beta})$
$\therefore$ the MAP estimator $\hat{\vec{\beta}}_{MAP}$ can be derived from

$$\begin{aligned}
\hat{\vec{\beta}}_{MAP} &= \arg\max_{\vec{\beta}} \textstyle\prod_{i=1}^{N} P(y_i, |\vec{x}_i, \vec{\beta})P(\vec{\beta}) \\
&= \arg\max_{\vec{\beta}} \textstyle\sum_{i=1}^{N}(-\frac{1}{2\sigma^2}(y_i - \vec{x}^T\vec{\beta})^2) - \frac{\lambda}{\sigma^2}\vec{\beta}^T\vec{\beta} \\
&= \arg\min_{\vec{\beta}} \tfrac{1}{2}\textstyle\sum_{i=1}^{N}(y_i - \vec{x}^T\vec{\beta})^2 + \lambda\vec{\beta}^T\vec{\beta} \\
&= \arg\min_{\vec{\beta}} \tfrac{1}{2}||\vec{Y} - X\vec{\beta}||_2^2 + \lambda||\vec{\beta}||_2^2
\end{aligned}$$

$\therefore$ The MAP estimator is equivalent to the solution of the following ridge regression problem:

$$J^*(\vec{\beta}) = \min_{\vec{\beta}} \frac{1}{2}||\vec{Y} - X\vec{\beta}||_2^2 + \lambda||\vec{\beta}||_2^2$$

4. Refer to the closed form solutions of (1) and (2) in the lecture slides, what might be an issue of $\hat{\vec{\beta}}_{MLE}$ if $d >> N$? How can $\hat{\vec{\beta}}_{MAP}$ possibly address it?
Answer:
From the lecture, we know the closed form solutions of (1) and (2) as below:

- $\hat{\vec{\beta}}_{MLE} = (X^T X)^{-1} X^T Y$
- $\hat{\vec{\beta}}_{MAP} = (X^T X + \lambda I)^{-1} X^T Y$

$\because$ X is a $N \times d$ matrix.
$\therefore$ $X^T X$ is a $d \times d$ matrix.
$\because$ $rank(X^T X) \leq rank(X) \leq min(d, N) = N << d$
$\therefore$ $X^T X$ is not invertible.
$\therefore$ $\hat{\vec{\beta}}_{MLE}$ is not feasible.
$\because$ $(X^T X + \lambda I)$ is a positive definite matrix, if $\lambda > 0$
$\therefore$ $(X^T X + \lambda I)$ is invertible.
$\therefore$ $\hat{\vec{\beta}}_{MAP}$ is feasible.

Part B, Problem 3: MLE, MAP and Logistic Regression

We learnt about Maximum Likelihood estimation in class. For a fixed set of data and underlying statistical model, the method of maximum likelihood selects the set of values of the model parameters that maximizes the likelihood function.

In this problem, we will look at two different ways of estimating parameters in a probability distribution. Suppose we observe n i.i.d. random variables $X_1, ..., X_n$, drawn from a distribution with parameter θ. That is, for each X_i and a natural number k,

$$P(X_i = k) = (1-\theta)^k \theta$$

Given some observed values of X_1 to X_n , we want to estimate the value of θ.

3.1 Maximum Likelihood Estimation

The first kind of estimator for θ we will consider is the Maximum Likelihood Estimator (MLE). The probability of observing given data is called the likelihood of the data, and the function that gives the likelihood for a given parameter $\hat{\theta}$ (which may or may not be equal to the true parameter θ) is called the likelihood function, written as $L(\hat{\theta})$. When we use MLE, we estimate θ by choosing the $\hat{\theta}$ that maximizes the likelihood.

$$\hat{\theta}_{MLE} = \arg\max_{\hat{\theta}} L(\hat{\theta})$$

It is often convenient to deal with the log-likelihood ($\ell(\hat{\theta}) = \log L(\hat{\theta})$) instead, and since log is an increasing function, the argmax also applies in the log space:

$$\hat{\theta}_{MLE} = \arg\max_{\hat{\theta}} \ell(\hat{\theta})$$

1. Given a dataset $\mathcal{D}$, containing observations $\{X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n\}$, write an expression for $\ell(\hat{\theta})$ as a function of $\mathcal{D}$ and $\hat{\theta}$. How does the order of the variables affect the function?
   Answer:
   $\because P(X_i = k_i|\theta) = (1-\theta)^{k_i}\theta$
   $\therefore$
   $$\begin{aligned} \ell(\hat{\theta}) &= \textstyle\prod_{i=1}^n P(X_i = k_i|\hat{\theta}) \\ &= (1-\theta)^{\sum_{i=1}^n k_i}\theta^n \end{aligned}$$

2. Derive an expression for the maximum likelihood estimate.
   Answer:
   Assume $K = \sum_{i=1}^n k_i$
   $$\begin{aligned} \hat{\theta}_{MLE} &= \arg\max_{\hat{\theta}} \ell(\hat{\theta}) \\ \left.\frac{d\ell(\hat{\theta})}{d\hat{\theta}}\right|_{\hat{\theta}_{MLE}} &= 0 \\ \left.((1-\theta)^{K-1}\theta^{n-1}(n-(n+K)\theta))\right|_{\hat{\theta}_{MLE}} &= 0 \\ \hat{\theta}_{MLE} &= \frac{n}{n+K} = \frac{n}{n+\sum_{i=1}^n k_i} \end{aligned}$$

3.2 Maximum a Posteriori Estimation

Now we assume that we have some prior knowledge about the true parameter θ. We express it by treating θ itself as a random variable and defining a prior probability distribution over it. Precisely, we suppose that the data $X_1, \ldots, X_n$ are drawn as follows:

- θ is drawn from the prior probability distribution

- Then $X_1, \dots, X_n$ are drawn independently from a Geometric distribution with θ as the parameter.

Now both X_i and θ are random variables, and they have a joint probability distribution. We now estimate θ as follows

$$\hat{\theta}_{MAP} = \arg\max_{\hat{\theta}} P(\theta = \hat{\theta}|X_1, \dots, X_n)$$

This is called Maximum a Posteriori (MAP) estimation. Using Bayes rule, we can rewrite the posterior probability as follows.

$$P(\theta = \hat{\theta}|X_i, \dots, X_n) = \frac{P(X_i, \dots, X_n|\theta = \hat{\theta})P(\theta = \hat{\theta})}{P(X_1, \dots, X_n)}$$

Applying this to the MAP estimate, we get the following expression. Notice that we can ignore the denominator since it is not a function of $\hat{\theta}$

$$\begin{aligned}
\hat{\theta}_{MAP} &= \arg\max_{\hat{\theta}} P(X_1, \dots, X_n|\theta = \hat{\theta})P(\theta = \hat{\theta}) \\
&= \arg\max_{\hat{\theta}} L(\hat{\theta})P(\theta = \hat{\theta}) \\
&= \arg\max_{\hat{\theta}} (\ell(\hat{\theta}) + \log P(\theta = \hat{\theta}))
\end{aligned}$$

Thus, the MAP estimator maximizes the sum of the log-likelihood and the log-probability of the prior distribution on θ. WHen the prior is a continuous distribution with density function p, we have

$$\hat{\theta}_{MAP} = \arg\max_{\hat{\theta}} (\ell(\hat{\theta}) + \log p(\hat{\theta}))$$

For this problem, we will use the Beta distribution (a popular choice when the data dstribution is Geometric or Bernoulli) as the prior, and the density function is given by

$$p(\hat{\theta}) = \frac{\hat{\theta}^{\alpha-1}(1-\hat{\theta})^{\beta-1}}{B(\alpha, \beta)}$$

where $B(\alpha, \beta)$ is the beta function.

4. Derive a close form expression for the maximum a posteriori estimate. (hint: If x^* maximizes f, $f'(x^*) = 0$).
   Answer:
   $\because P(X_i = k_i|\theta) = (1-\theta)^{k_i}\theta$ and $p(\hat{\theta}) = \frac{\hat{\theta}^{\alpha-1}(1-\hat{\theta})^{\beta-1}}{B(\alpha,\beta)}$
   $\therefore$ The MAP estimate of $\hat{\theta}$ can be derived by (assume $K = \sum_{i=1}^{n} k_i$):

$$\begin{aligned}
\hat{\theta}_{MAP} &= \arg\max_{\hat{\theta}} \textstyle\prod_{i=1}^{n} p(X_i = k_i|\hat{\theta})p(\hat{\theta}) \\
&= \arg\max_{\hat{\theta}} (1-\hat{\theta})^K \hat{\theta}^n p(\hat{\theta}) \\
&= \arg\max_{\hat{\theta}} K\ln(1-\hat{\theta}) + n\ln\hat{\theta} + \ln p(\hat{\theta}) \\
&= \arg\max_{\hat{\theta}} K\ln(1-\hat{\theta}) + n\ln\hat{\theta} + (\alpha-1)\ln\hat{\theta} + (\beta-1)\ln(1-\hat{\theta}) \\
&= \arg\max_{\hat{\theta}} (n+\alpha-1)\ln\hat{\theta} + (K+\beta-1)\ln(1-\hat{\theta})
\end{aligned}$$

   Assume the objective function $J(\hat{\theta}) = (n+\alpha-1)\ln\hat{\theta} + (K+\beta-1)\ln(1-\hat{\theta})$, then

$$\begin{aligned}
\left.\frac{dJ(\hat{\theta})}{d\hat{\theta}}\right|_{\hat{\theta}_{MAP}} &= 0 \\
\left.\left(\frac{n+\alpha-1}{\hat{\theta}} - \frac{K+\beta-1}{1-\hat{\theta}}\right)\right|_{\hat{\theta}_{MAP}} &= 0 \\
\hat{\theta}_{MAP} &= \frac{n+\alpha-1}{K+n+\alpha+\beta-2}
\end{aligned}$$

5. Is the bias of Maximum Likelihood Estimate (MLE) typically greater than or equal to the bias of Maximum A Posteriori (MAP) estimate? (Explain your answer in a sentence)
   Answer:
   No, it depends on the difference between the prior distribution and real distribution of parameters.

6. What can you say about the value of Maximum Likelihood Estimate (MLE) as compared to the value of Maximum A Posteriori (MAP) estimate with a uniform prior? Why?
Answer:
If the prior is a uniform distribution, then with such a weak priori, the MLE is same with MAP. Because the priori has no effect on the MAP estimation, and only the shared likelihood between both estimators will determine the estimation value.

3.3 Logistic Regression

In class, we wil learn about MLE of parametrers in logistic regression. For a given data $\vec{x} \in \mathbb{R}^p$, the probability of Y being 1 in logistic regression is

$$P(Y = 1|\vec{X} = \vec{x}) = \frac{\exp(w_0 + \vec{x}^T\vec{w})}{1 + \exp(w_0 + \vec{x}^T\vec{w})} \tag{3}$$

where w_0 and $\vec{w} = (w_1, w_2, \ldots, w_p)^T$ are model parameters. In this problem, we consider the maximum a posteriori setting, where we put a Gaussian prior on the parameters:

$$w_i \sim \mathcal{N}(\mu, 1),\ for\ i = 0, 1, 2, \ldots, p.$$

7. Choose a conjugate prior for Gaussian on μ (choose any higher parameters as you want to ease the computation). Assuming you are given a dataset with n training examples and p features, write down a formula for the conditional log posterior likelihood of the training data in terms of the class labels $y^{(i)}$, the features $x_1^{(i)}, \ldots, x_p^{(i)}$, and the parameters $w_0, w_1, \ldots, w_p$, where the superscript (i) denotes the sample index. This will be your objective function for gradient ascent.
Answer:
Choose $\mu \sim \mathcal{N}(\tilde{\mu}, \sigma^2)$ as a conjugate prior for Gaussian on μ, then the conditional log posterior likelihood of the training data is (assume $\vec{\mathcal{Y}} = [y^{(1)}, y^{(2)}, \ldots, y^{(n)}]$, and $\mathcal{X} = [\vec{x}^{(1)}, \vec{x}^{(2)}, \ldots, \vec{x}^{(n)}]^T$):

$$\begin{aligned} f(w_0, \vec{w}, \mu) &= \ln(P(\vec{\mathcal{Y}}|\mathcal{X}, w_0, \vec{w}, \mu)P(w_0, \vec{w}|\mu)P(\mu)) \\ &= \textstyle\sum_{i=1}^n \ln(P(y^{(i)}|\vec{x}^{(i)}, w_0, \vec{w}, \mu)) + \sum_{j=0}^p \ln(P(w_j|\mu)) + \ln(P(\mu)) \\ &= \textstyle\sum_{i=1}^n (y^{(i)}(w_0 + \sum_{k=1}^p w_k x_k^{(i)}) - \ln(1 + \exp(w_0 + \sum_{k=1}^p w_k x_k^{(i)}))) \\ &\quad -\tfrac{1}{2}\textstyle\sum_{j=0}^p (w_j - \mu)^2 - \frac{1}{2\sigma^2}(\mu - \tilde{\mu})^2 + C \end{aligned}$$

8. Compute the partial derivative of the objective with respect to w_0, to an arbitrary w_i, and μ, i.e. derive $\partial f/\partial w_0$, $\partial f/\partial w_i$, $\partial f/\partial \mu$ where f is the objective that you provided above. Use (3) to simplify the formula. What is the MAP estimation of μ given w_0 and $\vec{w}$
Answer:

- $\partial f/\partial w_0$:

$$\begin{aligned} \partial f/\partial w_0 &= \textstyle\sum_{i=1}^n (y^{(i)} - \frac{\exp(w_0 + \sum_{k=1}^p w_k x_k^{(i)})}{1 + \exp(w_0 + \sum_{k=1}^p w_k x_k^{(i)})}) \\ &= \textstyle\sum_{i=1}^n (y^{(i)} - P(Y^{(i)} = 1|\vec{x}^{(i)}, w_0, \vec{w})) \end{aligned}$$

- $\partial f/\partial w_i$:

$$\begin{aligned} \partial f/\partial w_i &= \textstyle\sum_{j=1}^n (y^{(j)} x_i^{(j)} - \frac{x_i^{(j)} \exp(w_0 + \sum_{k=1}^p w_k x_k^{(i)})}{1 + \exp(w_0 + \sum_{k=1}^p w_k x_k^{(j)})}) \\ &= \textstyle\sum_{j=1}^n x_i^{(j)} (y^{(j)} - P(Y^{(j)} = 1|\vec{x}^{(j)}, w_0, \vec{w})) \end{aligned}$$

- $\partial f/\partial \mu$:

$$\begin{aligned} \partial f/\partial \mu &= \textstyle\sum_{j=0}^p (w_j - \mu) - \frac{1}{\sigma^2}(\mu - \tilde{\mu}) \\ \partial f/\partial \mu\big|_{\hat{\mu}_{MAP}} &= 0 \\ \textstyle\sum_{j=0}^p (w_j - \mu) - \frac{1}{\sigma^2}(\mu - \tilde{\mu})\Big|_{\hat{\mu}_{MAP}} &= 0 \\ \hat{\mu}_{MAP} &= \frac{\sum_{j=0}^p w_j + \tilde{\mu}/\sigma^2}{p + 1 + 1/\sigma^2} \end{aligned}$$

Part C: Programming Exercise