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%

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

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

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:

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,

$$Bias[\hat{\vec{\theta}}] = \mathbb{E}[\hat{\vec{\theta}}] - \vec{\theta} \; and \; Var[\hat{\vec{\theta}}] = \mathbb{E}[(\hat{\vec{\theta}} - \vec{\theta})(\hat{\vec{\theta}} - \vec{\theta})]$$

1. Derive the maximum likelihood estimator:

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

Part B, Problem 2: Linear Regression

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

Part C: Programming Exercise