Machine Learning (CS 6140)
Homework 1

Instructor: Ehsan Elhamifar
Due Date: October 13, 2016, 11:45am

1) Probability and Random Variables: State true or false. Here Ω denotes the sample space and A^c denotes the complement of the event A.

1. Assume $P(B) > 0$, then $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.
2. For any $A, B \subseteq \Omega$ such that $P(B) > 0$, $P(A^c) > 0$, $P(A|B) + P(B|A^c) = 1$.
3. For any $A, B \subseteq \Omega$ such that $0 < P(B) < 1$, $P(A|B) + P(A|B^c) = 1$.
4. For any $A, B \subseteq \Omega$, $P(B^c \cup (A \cap B)) + (B \cap (A \cup A^c)) = 1$.
5. Let $\{A_i\}_{i=1}^n$ be mutually independent. Then, P$(\cap_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$.

2) Discrete and Continuous Distributions: Write down the formula of the probability density/mass functions of random variable X.

1. 1-d Gaussian distribution, $X \sim N(x; \mu, \sigma^2)$.
2. Bernoulli distribution, $X \sim$ Bernoulli$(p), 0 < p < 1$.
3. Uniform distribution, $X \sim$ Unif$(a, b), a < b$.
4. Exponential distribution, $X \sim$ Exp$(\lambda), \lambda > 0$.
5. Poisson distribution, $X \sim$ Poisson$(\lambda), \lambda > 0$.

3) Vector Norms: Draw the regions corresponding to vectors $x \in \mathbb{R}^2$ with the following norms:

1. $\|x\|_1 \leq 1$ (Recall that $\|x\|_1 = \sum_i |x_i|$
2. $\|x\|_2 \leq 1$ (Recall that $\|x\|_2 = \sqrt{\sum_i x_i^2}$)
3. $\|x\|_\infty \leq 1$ (Recall that $|x\|_\infty = \text{max}_i |x_i|$)

4) Geometry: Prove that these are true or false. Provide all steps.

1. The Euclidean distance from the origin to the hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$ is $\frac{|b|}{\|\mathbf{w}\|_2}$.
2. The Euclidean distance between two parallel hyperplane $\mathbf{w}^T\mathbf{x}+b_1 = 0$ and $\mathbf{w}^T\mathbf{x}+b_2 = \frac{|b_1-b_2|}{\|\mathbf{w}\|_2}$ (Hint: you can use the result from the last question to help you prove this one).

5) Multi-output linear regression: When we have multiple independent outputs in linear regression, the model becomes

$$p(\mathbf{y}|\mathbf{x}, \mathbf{W}) = \prod_{j=1}^{M} N(y_i|\mathbf{w}_j^T \mathbf{x}_i, \sigma_j^2) \tag{7.89}$$

Since the likelihood factorizes across dimensions, so does the MLE. Thus

$$\hat{\mathbf{W}} = [\hat{\mathbf{w}}_1, \cdots, \hat{\mathbf{w}}_M] \tag{7.90}$$

where $\hat{\mathbf{w}}_j = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{Y}_{:,j}$.

In this exercise we apply this result to a model with 2 dimensional response vector $y_i \in \mathbb{R}^2$. Suppose we have some binary input data, $x_i \in \{0, 1\}$. The training data is as follows:

x	y
0	$(-1,-1)^T$
0	$(-1,-2)^T$
0	$(-2,-1)^T$
1	$(1,1)^T$
1	$(1,2)^T$
1	$(2,1)^T$

Let us embed each x_i into $2d$ using the following basis function:

$$\phi(0) = (1, 0)^T, \phi(1) = (0, 1)^T \tag{7.91}$$

The model becomes

$$\hat{y} = \mathbf{W}^T \phi(x) \tag{7.92}$$

where $\mathbf{W}$ is a 2×2 matrix. Compute the MLE for $\mathbf{W}$ from the above data.

6) Centering and ridge regression: Assume that $\overline{x} = 0$, so that the input data has been centered. Show that the optimizer of

$$J(\mathbf{w}, w_0) = (\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1})^T(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1}) + \lambda\mathbf{w}^T\mathbf{w} \tag{7.93}$$

is

$$\hat{w}_0 = \overline{y} \tag{7.94}$$

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \tag{7.95}$$

7) MAP estimation for the Bernoulli with non-conjugate priors: In the book, we discussed Bayesian inference of a Bernoulli rate parameter with the prior $p(\theta) = \text{Beta}(\theta|\alpha, \beta)$. We know that, with this prior, the MAP estimate is given by:

$$\hat{\theta} = \frac{N_1 + \alpha - 1}{N + \alpha + \beta - 2} \tag{3.100}$$

where N_1 is the number of heads, N_0 is the number of tails, and $N = N_0 + N_1$ is the total number of trials.

1. Now consider the following prior, that believes the coin is fair, or is slightly biased towards tails:
$$p(\theta) = \begin{cases} 0.5 & \text{if } \theta = 0.5 \\ 0.5 & \text{if } \theta = 0.4 \\ 0 & \text{otherwise} \end{cases} \tag{3.101}$$
Derive the MAP estimate under this prior as a function of N_1 and N.

2. Suppose the true parameters is $\theta = 0.41$. Which prior leads to a better estimate when N is small? Which prior leads to a better estimate when N is large?

8) Gaussian Discriminant Analysis: The multivariate normal distribution in n-dimensions, also called the multi-variate Gaussian distribution, is parameterized by a mean vector $\mu \in \mathbb{R}^n$ and a covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$, where $\Sigma \geq 0$ is symmetric and positive semi-definite. Also written "$N(\mu, \Sigma)$", its density is given by:

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right). \tag{1}$$

In the equation above, "$|\Sigma|$" denotes the determinant of the matrix Σ

When we have a classification problem in which the input feature x are continuous-valued random variables, we can then use the Gaussian Discriminant Analysis (GDA) model, which models $p(x|y)$ using a multivariate normal distribution. The model is:

$$y \sim \text{Bernoulli}(\phi)$$
$$x|y = 0 \sim N(\mu_0, \Sigma)$$
$$x|y = 1 \sim N(\mu_1, \Sigma)$$

Given a training dataset $\{(\boldsymbol{x}^1, y^1), \ldots, (\boldsymbol{x}^N, y^N)\}$, write down the likelihood (log-likelihood) and derive MLE estimates for the means μ_0, μ_1 and covariance Σ of the GDA.

9) Linear Regression Implementation: A) Write down a code in Python whose input is a training dataset $\{(\boldsymbol{x}^1, y^1), \ldots, (\boldsymbol{x}^N, y^N)\}$ and its output is the weight vector $\boldsymbol{w}$ in the linear regression model $y = \boldsymbol{w}^\top \phi(\boldsymbol{x})$, for a given nonlinear mapping $\phi(\cdot)$ Consider n-degree polynomials $\phi(\cdot) = \begin{bmatrix} 1 & x & x^2 & \cdots & x^n \end{bmatrix}$. B) Download the dataset on the course webpage. Run the code on the training dataset to compute $\boldsymbol{w}$ and evaluate on the test dataset. Report $\boldsymbol{w}$, training error and test error. C) Write a code that applies Ridge regression to the dataset to compute $\boldsymbol{w}$ for a given λ. Use a K-fold cross validation on the training dataset to obtain the best regularization λ and apply the result to the test data. Report the optimal λ, $\boldsymbol{w}$, test and training set errors for $K \in \{2, 5, 10, N\}$. In all cases try $n = \{2, 5, 10, 20\}$.

10) Logistic Regression Implementation: A) Write down a code in Python whose input is a training dataset $\{(\boldsymbol{x}^1, y^1), \ldots, (\boldsymbol{x}^N, y^N)\}$ and its output is the weight vector $\boldsymbol{w}$ in the logistic regression model $y = \sigma(\boldsymbol{w}^\top \phi(\boldsymbol{x}))$, for a given nonlinear mapping $\phi(\cdot)$. B) Download the dataset on the course webpage. Run the code on the training dataset to compute $\boldsymbol{w}$ and evaluate on the test dataset. Report $\boldsymbol{w}$, training and test set classification errors.