Machine Learning (CS 6140)
Homework 1

Instructor: Ehsan Elhamifar
Due Date: October 12, 2017, 11:45am

1) Probability and Random Variables: State true or false. If true, prove it. If false, either prove or demonstrate by a counter example. Here Ω denotes the sample space and A^c denotes the complement of the event A.

1. For any $A, B \subseteq \Omega$ such that $P(A) > 0$, $P(B^c) > 0$, $P(B|A) + P(A|B^c) = 1$.
2. For any $A, B \subseteq \Omega$ such that $0 < P(B) < 1$, $P(A|B) + P(A|B^c) = 1$.
3. For any $A, B \subseteq \Omega$, $P(B^c \cup (A \cap B)) + P(A^c \cap B) = 1$.
4. Let $\{A_i\}_{i=1}^n$ be mutually independent. Then, $P(\cup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$.
5. Let $\{(A_i, B_i)\}_{i=1}^n$ be mutually independent, i.e., (A_i, B_i) is independent from (A_j, B_j) for every $i \neq j$. Then $P(A_1, \ldots, A_n|B_1, \ldots, B_n) = \prod_{i=1}^n P(A_i|B_i)$

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

1. k-dimensional Gaussian distribution (multi-variate Gaussian), $X \sim \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$.
2. Laplace distribution with mean μ and variance $2\sigma^2$.
3. Bernoulli distribution, $X \sim \text{Bernoulli}(p), 0 < p < 1$.
4. Multinomial distribution with N trials and L outcomes with probabilities $\theta_1, \ldots, \theta_L$.
5. Dirichlet distribution of order L with parameters $\alpha_1, \ldots, \alpha_L$.
6. Uniform distribution, $X \sim \text{Unif}(a, b), a < b$.
7. Exponential distribution, $X \sim \text{Exp}(\lambda), \lambda > 0$.
8. Poisson distribution, $X \sim \text{Poisson}(\lambda), \lambda > 0$.

3) Positive-Definite Matrices: A symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{m \times m}$ is positive-semidefinite if $\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x} \geq 0$ for every $\boldsymbol{x} \in \mathbb{R}^m$, where $\boldsymbol{x} \neq \boldsymbol{0}$. Equivalently, $\boldsymbol{A}$ is positive-semidefinite if all eigenvalues of $\boldsymbol{A}$ are non-negative. Prove or disprove (by a counter example) that the following matrices are positive-semidefinite.

1. $\boldsymbol{A} = \boldsymbol{B}^\top \boldsymbol{B}$ for an arbitrary $\boldsymbol{B} \in \mathbb{R}^{m \times n}$
2. $\boldsymbol{A} = \begin{bmatrix} 8 & -5 & -3 \\ -5 & 5 & 0 \\ -3 & 0 & 3 \end{bmatrix}$

3. $\boldsymbol{A} = \boldsymbol{B} + \boldsymbol{B}^\top + \boldsymbol{B}^\top \boldsymbol{B}$ for an arbitrary $\boldsymbol{B} \in \mathbb{R}^{n \times n}$

4) Convexity of Linear Regression: In the class, we studied several models for linear regression. Let $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ and $\boldsymbol{Y} \in \mathbb{R}^N$ denote matrices of input features and outputs/responses, respectively. Let $\boldsymbol{\theta} \in \mathbb{R}^d$ denote the vector of unknown parameters.

a) Show that the following objective functions for linear regression are convex with respect to $\boldsymbol{\theta}$.

1. Vanilla/Basic regression: $J_1(\boldsymbol{\theta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2$
2. Ridge regression: $J_2(\boldsymbol{\theta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2$
3. Lasso regression: $J_3(\boldsymbol{\theta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1$

b) What conditions do we need to impose on $\boldsymbol{X}$ and/or $\boldsymbol{Y}$ in each of the above cases, so that the solution for $\boldsymbol{\theta}$ be unique?

5) Regression using Huber Loss: In the class, we defined the Huber loss as

$$\ell_\delta(e) = \begin{cases} \frac{1}{2}e^2 & \text{if } |e| \le \delta \\ \delta|e| - \frac{1}{2}\delta^2 & \text{if } |e| > \delta \end{cases}$$

Consider the robust regression model

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \ell_\delta(y_i - \boldsymbol{\theta}^\top \boldsymbol{x}_i),$$

where $\boldsymbol{x}_i$ and y_i denote the i-th input sample and output/response, respectively and $\boldsymbol{\theta}$ is the unknown parameter vector.

a) Provide the steps of the batch gradient descent in order to obtain the solution for $\boldsymbol{\theta}$.

b) Provide the steps of the stochastic gradient descent using mini-batches of size 1, i.e., one sample in each mini-batch, in order to obtain the solution for $\boldsymbol{\theta}$.

6) PAC Confidence Bounds: In the class, we studied the problem of maximum likelihood estimation of a Bernoulli random variable (taking values in $\{0, 1\}$), where the true probability of being 1 is assumed to be θ^o. We showed that using the maximum likelihood estimation on a dataset with N samples, the ML estimate is given by $\hat{\theta} = \sum_{i=1}^N x_i/N$. Moreover, we showed that

$$P\left(|\hat{\theta} - \theta^o| \ge \epsilon\right) \le 2e^{-N\epsilon^2}.$$

Consider the example of flipping a coin N times with the true probability of 'Head' to be θ^o. How many trials (flipping the coin) we need to have in order to be confident that with probability at least 0.95, the estimate of the maximum likelihood for the probability of 'Head' will be within 0.1 distance of the true value?

7) Probabilistic Regression with Prior on Parameters: Consider the probabilistic model of regression, where $p(y|\boldsymbol{x}, \boldsymbol{\theta})$ is a Normal distribution with mean $\theta^\top \boldsymbol{x}$ and variance σ^2, i.e., $\mathcal{N}(\boldsymbol{\theta}^\top \boldsymbol{x}, \sigma^2)$.

We would like to determine $\boldsymbol{\theta}$ using a dataset of N samples $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$. Assume we have prior information about the distribution of θ and our goal is to determine the Maximum A Posteriori (MAP) estimate of $\boldsymbol{\theta}$ using the dataset and the prior information. For each of the following cases, provide the optimization from which we can obtain the MAP solution.

1. $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, 1/\lambda \boldsymbol{I})$, where $\boldsymbol{I}$ denotes the identity matrix.

2. Each element of $\boldsymbol{\theta}$ has a Laplace distribution with mean 0 and variance $2/\lambda^2$.

8) MAP estimation for the Bernoulli with non-conjugate priors: Consider a Bernoulli random variable x with $p(x = 1) = \theta$. In the class, we discussed MAP estimation of the Bernoulli rate parameter θ with the prior $p(\theta) = \text{Beta}(\theta|\alpha, \beta)$. We know that, with this prior, the MAP estimate is given by:

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

where N_1 is the number of trails where $x_i = 1$ (e.g., heads in flipping a coin), N_0 is the number of trials where $x_i = 0$ (e.g., tails in flipping a coin) 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 heads:

$$p(\theta) = \begin{cases} 0.5 & \text{if } \theta = 0.5 \\ 0.5 & \text{if } \theta = 0.6 \\ 0 & \text{otherwise} \end{cases}$$

   Derive the MAP estimate under this prior as a function of N_1 and N.

2. Suppose the true parameters is $\theta = 0.61$. Which prior leads to a better estimate when N is small? Which prior leads to a better estimate when N is large?

9) Gaussian Naive Bayes: The multivariate normal distribution in k-dimensions, also called the multi-variate Gaussian distribution and denoted by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, is parameterized by a mean vector $\boldsymbol{\mu} \in \mathbb{R}^k$ and a covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{k \times k}$, where $\boldsymbol{\Sigma} \geq \mathbf{0}$ is positive semi-definite.
Consisder a classification problem in which the input feature $\boldsymbol{x} \in \mathbb{R}^k$ are continuous-valued random variables, we can then use the Gaussian Naive Bayes (GNB) model, which models $p(\boldsymbol{x}|y)$ using a multivariate normal distribution. The model is given by

$$y \sim \text{Bernoulli}(\phi)$$

$$\boldsymbol{x}|y = 0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$$
$$\boldsymbol{x}|y = 1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$$

where we assume that $\boldsymbol{\Sigma}$ is a diagonal matrix, $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_k^2)$. 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 $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, covariance $\boldsymbol{\Sigma}$ and the class prior ϕ of the GNB.

10) 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 $\boldsymbol{\theta}$ in the linear regression model $y = \boldsymbol{\theta}^\top \phi(\boldsymbol{x})$, for a given nonlinear

mapping $\phi(\cdot)$. Implement two cases: i) using the closed-form solution, ii) using a stochastic gradient descent on mini-batches of size m.

b) Consider n-degree polynomials, $\phi(\cdot) = \begin{bmatrix} 1 & x & x^2 & \cdots & x^n \end{bmatrix}$. Download the dataset on the course webpage and work with 'dataset1'. Run the code on the training data to compute $\boldsymbol{\theta}$ for $n \in \{2, 3, 5\}$. Evaluate the regression error on both training and the test data. Report $\boldsymbol{\theta}$, training error and test error for both implementation (closed-form vs gradient descent). What is the effect of the size of the mini-batch on the speed and testing error of the solution.

c) Download the dataset on the course webpage and work with 'dataset2'. Write a code in Python that applies Ridge regression to the dataset to compute $\boldsymbol{\theta}$ for a given λ. Implement two cases: using a closed-form solution and using a stochastic gradient descent method with mini-batches of size m. Use K-fold cross validation on the training dataset to obtain the best regularization λ and apply the optimal $\boldsymbol{\theta}$ to compute the regression error on test samples. Report the optimal λ, $\boldsymbol{\theta}$, test and training set errors for $K \in \{2, 10, N\}$, where N is the number of samples. In all cases try $n \in \{2, 3, 5\}$. How does the test error change as a function of λ and n?