Machine Learning 10-701 January 20, 2015
http://alex.smola.org/teaching/10-701-15 due on January 27, 2015
Carnegie Mellon University

Homework 1

FINAL VERSION

Before this final version on January 20th, there was a version uploaded on January 19th. To avoid confusion, here we list the changes made from the previous version.

- Score distribution has been changed. The course staff decided that the problem 2 is much harder than problem 1, and downgraded the point of problem 1 to 30 points.
- Previous section 1.1 Expectation has been removed and now section 1.1 is Independence.
- Some more explanations have been added to problem 2.
- The deadline has been extended for a day from January 26th to January 27th.
- Lastly, the following instructions on next section were clarified.

INSTRUCTIONS

- The homework is due at 9:00am on January 27, 2015. Anything that is received after that time will be considered to be late.
- Answers to every theory questions need to be submitted electronically on Autolab. Only PDF is acceptable (e.g. LaTeX or handwritten and scanned).
- Make sure you prepare the answers to each question separately. This helps us dispatch the problems to different graders.
- Collaboration on solving the homework is allowed. Discussions are encouraged but you should think about the problems on your own.
- When you do collaborate, you should list your collaborators! Also cite your resources, in case you got some inspiration from other resources (books, websites, papers).
- If you do collaborate with someone or use a book or website, you are expected to write up your solution independently. That is, close the book and all of your notes before starting to write up your solution.

1 Probability Review and Bayesian Spam Filter [Zhou; 30pts]

Probability is, in many ways, the most fundamental mathematical technique for machine learning. This problem will review several basic notions from probability and make sure that you remember how to do some elementary proofs.

Recall that for a discrete random variable X whose values are integers, we frequently use the notation $P(X = x)$. If a random variable Y is continuous, we typically use a "density function" $f(Y = y)$. The conditions for $P(X = x)$ to be a valid probability distribution are that $\Sigma_{-\infty}^{\infty} P(X = x) = 1$ and $P(X = x) \geq 0$. Similarly for $f(Y = y)$ to be a valid continuous distribution, $\int_{-\infty}^{\infty} f(Y = y)dy = 1$ and $f(Y = y) \geq 0$.

Sometimes the underlying probability space has more than one variable (for example, the height and weight of a person). In this case, we may use notation like $f(X = x; Y = y)$ to denote the probability density function in several dimensions.

1.1 Independence [5 pts]

Intuitively, two random variables X and Y are "independent" if knowledge of the value of one tells you nothing at all about the value of the other. Precisely, if X and Y are discrete, independence means that $P(X = x; Y = y) = P(X = x)P(Y = y)$, and if they are continuous, $f(X = x; Y = y) = f(X = x)f(Y = y)$.

Show the following, for independent random variable X and Y:

$$E[XY] = E[X]E[Y] \text{ for discrete and continous cases respectively.}$$

1.2 Spam filtering equation [5 pts]

Naive Bayes classifiers work by correlating the use of tokens (typically words, or sometimes other things), with spam and non-spam e-mails and then using Bayesian inference to calculate a probability that an email is or is not spam. Naive Bayes spam filtering is a baseline technique for dealing with spam. Let's suppose the suspected message contains the word "replica". Most people who are used to receiving e-mail know that this message is likely to be spam, more precisely a proposal to sell counterfeit copies of well-known brands of watches. The spam detection software, however, does not "know" such facts; all it can do is compute probabilities.
Let $Pr(S|W)$ be the probability that a message is a spam given the word "replica" appears in it. Express $Pr(S|W)$ in terms of the following components we provide:

1. $Pr(S)$ is the overall probability that any given message is spam;
2. $Pr(W|S)$ is the probability that the word "replica" appears in spam messages;
3. $Pr(H)$ is the overall probability that any given message is not spam (is "ham");
4. $Pr(W|H)$ is the probability that the word "replica" appears in ham messages.

1.3 I.I.D. assumption in spam filters [10 pts]

Give at least 4 cases how "Independent and identically distributed random variables (i.i.d.)" assumption can be violated for spam filtering. Fore each case, please explain why i.i.d. is violated and how spammers might exploit the situation.

1.4 Poison the Bayesian spam filter [10 pts]

Depending on the implementation, Bayesian spam filtering may be susceptible to Bayesian poisoning, a technique used by spammers in an attempt to degrade the effectiveness of spam filters that rely on Bayesian filtering. Think of some ideas to make your spam pass the Bayesian spam filter based on the equation.

2 Regression [Jay-Yoon; 70pts]

The objective of this problem is to gain knowledge on Linear Regression, Maximum Likelihood Estimation (MLE), Maximum-a-Posteriori Estimation (MAP) and the variants of Regression problems with introduction of reularization terms.

2.1 Linear Regression: MLE and Least Squares [25 pts]

Consider a linear model with some Gaussian noise:

$$Y_i = \langle X_i, w \rangle + b + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2),\ i = 1, \ldots, n. \tag{1}$$

Where $Y_i \in \mathbb{R}$ is a scalar, $X_i \in \mathbb{R}^d$ is a d-dimensional vector, $b \in \mathbb{R}$ is a constant, $w \in \mathbb{R}^d$ is d-dimensional weight on X_i, and ϵ_i is a i.i.d. Gaussian noise with variance σ^2. Given the data $X_i, i = 1, \dots, n$, it is our goal to estimate w and b which specify the model.

We will show that solving the linear model (1) with MLE method is the same as solving the following Least Squares problem,

$$\operatorname*{argmin}_{\boldsymbol{\beta}} \, (\mathbf{y} - \mathbf{X}'\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}'\boldsymbol{\beta}), \tag{2}$$

where $Y = (Y_1, \dots, Y_n)^T$, $X_i' = (1, X_i)^T$, $\mathbf{X}' = (X_1', \dots, X_n')^T$ and $\boldsymbol{\beta} = (b, w)^T$.

1. From the model (1), derive the conditional distribution of $Y_i | X_i, w, b$. Again, X_i is a fixed data point.
2. Assuming i.i.d. between each ϵ_i, $i = 1, \dots, n$, give an explicit expression for the loglikelihood, $\log P(Y|\boldsymbol{\beta})$ of the data.

   Note: Notation Y and $\boldsymbol{\beta}$ is given on (2). $X_i,\ i = 1, \dots, n$ is fixed data points. Also given ϵ_is are i.i.d., so $P(Y|\boldsymbol{\beta}) = \prod_i P(Y_i|w, b)$.
3. Now show that solving for $\boldsymbol{\beta}$ that maximizes the loglikelihood, i.e. MLE, is the same as solving the Least Square problem of (2).
4. Derive $\boldsymbol{\beta}$ that maximizes the loglikelihood.

   Assume $\mathbf{X}'$ has full rank on column space.
   Hint: This link has some reference to matrix calculus: http://www.cs.nyu.edu/~roweis/notes/matrixid.pdf.

2.2 Nonlinear Regression and Regularization [15 pts]

Consider higher order term, $\phi(X_i) = (1, X_i, X_i^2 \dots, X_i^k)^T$, then we can model Y in kth-order model as following:

$$Y_i = \langle \phi(X_i), \boldsymbol{\beta'} \rangle + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2),\ i = 1, \dots, n. \tag{3}$$

where all the definitions are from equation (1) except $\boldsymbol{\beta'} \in \mathbb{R}^{k+1}$ and for simplicity let $d = 1$ for $X_i \in \mathbb{R}^d$.

1. As we did in section 2.1, show that optimal $\boldsymbol{\beta'}$ of (3) is $\left(\phi(\mathbf{X})^T\phi(\mathbf{X})\right)^{-1}\phi(\mathbf{X})^T\mathbf{Y}$ if we use MLE.

   Note: $\phi(\mathbf{X}) = (\phi(X_1), \phi(X_2), \dots, \phi(X_n))^T$ and assume $\phi(\mathbf{X})$ has full rank on column space.

   Hint: You are not expected to write the whole steps again. Foucs on the change from the loglikelihood expressions of 2.1 and derive the optimization problem.
2. In case $\phi(\mathbf{X})^T\phi(\mathbf{X})$ is not invertible, you can add diagnoal term $\lambda\mathbf{I}_{k+1}$ to it so that $\phi(\mathbf{X})^T\phi(\mathbf{X}) + \lambda\mathbf{I}_{k+1}$ becomes invertible with $\mathbf{I}_{k+1}$, an identity matrix of size n, and $\lambda > 0$. Show that $\left(\phi(\mathbf{X})^T\phi(\mathbf{X}) + \lambda\mathbf{I}_{k+1}\right)^{-1}\phi(\mathbf{X})^T\mathbf{Y}$ is the solution of the optimization problem,

$$\operatorname*{argmin}_{\boldsymbol{\beta'}} \|\mathbf{Y} - \phi(\mathbf{X})\boldsymbol{\beta'}\|_2^2 + \lambda\|\boldsymbol{\beta'}\|_2^2. \tag{4}$$

2.3 MAP [30 pts]

From (3), now consider a case where $\boldsymbol{\beta'}$ has a prior distribution $\boldsymbol{\beta'} \sim \mathcal{N}(0, \eta^2 \mathbf{I}_{k+1})$.

1. Write the posterior distribution of $\boldsymbol{\beta'}|Y_i$ given ith sample and $\boldsymbol{\beta'}|\mathbf{Y}$ given whole data, respectively. Assume independence between $\boldsymbol{\beta'}$ and noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2),\ i = 1, \ldots, n$.

   Hint1: Use Baye's rule and follow similar steps with 2.1. $\Pr(\boldsymbol{\beta'}|Y_i) = \frac{\Pr(\boldsymbol{\beta'}) \Pr(Y_i|\boldsymbol{\beta'})}{\Pr(Y_i)}$
   Hint2: Sum of two independent Normal variable follows Normal distribution. i.e. $X \sim \mathcal{N}(a, b^2), Y \sim \mathcal{N}(c, d^2), X \perp Y \rightarrow X + Y \sim \mathcal{N}(a + c, b^2 + d^2)$.

2. MAP estimate $\boldsymbol{\beta'}$ is defined as $\boldsymbol{\beta'}$ that establishes the mode of posterior $P(\boldsymbol{\beta'}|\mathbf{X}, \mathbf{Y})$. Show that solving for MAP estimate leads to the problem (4) if λ can be expressed in terms σ and η, i.e. $\lambda = g(\sigma, \eta)$. Find explicit expression for $g(\sigma, \eta)$.

3. Describe one potential problem in the absence of regularization term in (4) and how the regularziation term can alleviate the potential problem.

   Note: Let's skip the case of non-invertible matrix as we already covered the case in section 2.2.