# CS74/174, Winter 2013, Problem Set # 1

January 17, 2013

**Due: January 31, 11:59pm**

This problem set requires you to provide written answers to a few technical/theory questions. In addition, you will need to implement some learning algorithms in MATLAB and apply them to data sets enclosed with this homework. You can submit your answers to the technical questions either via Blackboard in electronic form (e.g. by typing your solutions in Latex or Word), or manually (handwrite your answers and drop them in the course mailbox located near the main entrance of Sudikoff). If you choose to hand in manually your answers and you submit late, please write the date and time of your submission on the homework. It is a violation of the honor code to report an incorrect submission date. The MATLAB code must be submitted via Blackboard: include all sources and the figures in a single folder, zip up the entire folder, and submit the compressed file. If you submit your technical answers electronically, include your document in the same folder containing the source code and the figures. Please use the following naming convention for the source code and figure files: the main script/function for question - say - 1(b) should be saved as file “q1b.m” and the corresponding output figure/plot as file “q1b.fig” (you can save .fig files in MATLAB using the command “saveas”).

1. **[10 points]** Using the definition $\mathrm{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right]$ show that $\mathrm{var}[f]$ satisfies the equation $\mathrm{var}[f] = \mathbb{E}\left[f(x)^2\right] - \mathbb{E}[f(x)]^2$.

2. **[10 points]** In this problem you will show that the sum of two Gaussian random variables is itself a Gaussian random variable. Let $\mathbf{x}$ and $\mathbf{z}$ be two multidimensional random vectors having Gaussian distributions $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mu_{\mathbf{x}}, \mathbf{\Sigma}_{\mathbf{x}})$ and $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mu_{\mathbf{z}}, \mathbf{\Sigma}_{\mathbf{z}})$ respectively. Let us define random vector $\mathbf{y}$ to be $\mathbf{y} = \mathbf{x} + \mathbf{z}$.

   (a) [5 points] Given these assumptions, write the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$, i.e., $p(\mathbf{y}|\mathbf{x})$. Please make sure to explain your answer (although no formal mathematical derivation is needed).

   (b) [5 points] Based on the result above, show that $\mathbf{y}$ is a Gaussian random variable and write its mean and covariance in terms of parameters $\{\mu_{\mathbf{x}}, \mathbf{\Sigma}_{\mathbf{x}}, \mu_{\mathbf{z}}, \mathbf{\Sigma}_{\mathbf{z}}\}$. In your proof, you can use the following theorem:

   *Theorem:* Let $\mathbf{u}, \mathbf{v}$ be two random vectors such that $p(\mathbf{u}) = \mathcal{N}(\mathbf{u}; \mu, \mathbf{\Sigma})$ and $p(\mathbf{v}|\mathbf{u}) = \mathcal{N}(\mathbf{v}; \mathbf{A}\mathbf{u} + \mathbf{b}, \mathbf{L})$ where $\mu, \mathbf{\Sigma}, \mathbf{A}, \mathbf{b}$ and $\mathbf{L}$ are known parameters. Then the marginal distribution of $\mathbf{v}$ is given by $p(\mathbf{v}) = \mathcal{N}(\mathbf{v}; \mathbf{A}\mu + \mathbf{b}, \mathbf{L} + \mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top)$.

3. **[15 points]** You meet Bob and you start talking about movies with him. You tell him about your two favorite ones: "A Beautiful Mind" ($M_1$) and "Schindler's List" ($M_2$). Bob tells you that since you like these two movies you might also like "Titanic" ($M_3$).

   Now you want to find out if it is truly worth watching $M_3$. You know the taste of your friends: 90% of them liked $M_1$ and 60% liked $M_2$. You also assume that these opinions are independent from each other. Also, by directing polling the friends that have already watched $M_3$ you discover that:

$$\begin{aligned} P(M_3 = l|M_1 = l, M_2 = l) &= 0.9 \\ P(M_3 = l|M_1 = l, M_2 = d) &= 0.7 \\ P(M_3 = l|M_1 = d, M_2 = l) &= 0.4 \\ P(M_3 = l|M_1 = d, M_2 = d) &= 0.2 \end{aligned}$$

   where for example $P(M_3 = l|M_1 = d, M_2 = l)$ indicates the fraction of your friends who liked $M_3$ among the subset that disliked $M_1$ and liked $M_2$.

   (a) [10 points] What is the probability that one of your friends likes $M_3$, i.e., $P(M_3 = l)$?

   (b) [5 points] Calculate $P(M_1 = l, M_2 = l|M_3 = l)$. What does this value tell you? Does Bob agree with your friends?

4. **[20 points]** Let's flip a coin. Let $c \in \{0, 1\}$ be a random variable indicating the result of the flip (1 for heads, 0 for tails). The probability that the coin lands heads on any trial is given by a parameter $\mu$. Note that we can write the distribution of $c$ as: $P(c\ ; \mu) = \mu^c(1 - \mu)^{1-c}$. We flip the coin $m$ times, and denote the result of the $i$-th flip by variable $c^{(i)}$. We assume that the coin flips are *independent*. We observe heads $H$ times.

   (a) [2 points] Write the likelihood function, i.e. the probability of the data $\mathcal{D} = \{c^{(1)}, ..., c^{(m)}\}$ given the model described above. Keep in mind that the likelihood is the probability of the observed training set $\mathcal{D}$ (rather than all possible training sets containing $H$ heads in $m$ examples). Thus your likelihood function should *not* include a combinatorial term counting the number of different ways $H$ heads could occur in $m$ trials.

   (b) [5 points] Derive the parameter $\mu$ using Maximum Likelihood estimation (hint: maximize the log likelihood).

   (c) [1 point] Using MATLAB, plot the likelihood function for the three cases: $\{m = 1, H = 1\}$, $\{m = 100, H = 100\}$, and $\{m = 100, H = 80\}$.

   Let us now assume that we have good reasons to believe that the coin is not counterfeit. However, we are not completely sure. We model this prior knowledge in the form of a prior distribution over $\mu$:

$$p(\mu; a) = \frac{1}{Z}\mu^{a-1}(1 - \mu)^{a-1} \tag{1}$$

   where $a$ is a parameter governing the prior distribution and $Z$ is a normalization constant (so that $\int_0^1 p(\mu; a)d\mu = 1$).

(d) [1 point] Plot in a single figure the prior for $a = 2$, and $a = 10$. The correct normalization constant to use is $Z = 1/6$ for $a = 2$, and $Z = 1/923780$ for $a = 10$.

(e) [6 points] Assuming the prior $p(\mu; a)$ in Eq. 1, derive the analytical expression of parameter $\mu$ using Maximum A Posteriori (MAP) estimation (hint: maximize the log posterior and drop the term related to the evidence, i.e. the denominator in Bayes' rule).

(f) [3 points] Can you provide a simple interpretation of parameter $a$ in this MAP estimation?

(g) [2 points] Plot the posterior (again, disregard the normalization factor related to the evidence) for $a = 10$, and the same flipping results considered above, i.e. $\{m = 1, H = 1\}$, $\{m = 100, H = 100\}$, and $\{m = 100, H = 80\}$.

5. **[30 points]** Write MATLAB code implementing a regression algorithm for multi-dimensional inputs $x$ and 1D outputs $y$. Your software must learn the regression hypothesis by minimizing the regularized least-square objective

$$E(\theta) = \frac{1}{2}\sum_{i=1}^{m}\left[y^{(i)} - \theta^T b(x^{(i)})\right]^2 + \frac{\lambda}{2}||\theta||^2 \tag{2}$$

for the following two distinct choices of the feature vector $b(x)$:
$b^l(x) = [1, x_1, ..., x_n]$ (where, as usual, $x_j$ denotes the $j$-th element of vector $x$), and
$b^q(x) = [1, x_1, ..., x_n, x_1^2, x_1x_2, ..., x_1x_n, x_2^2, x_2x_3, ..., x_2x_n, ...., x_n^2]$ (i.e. a quadratic function of the elements of $x$). Use the closed-form solution given in class to compute the parameter vector $\theta$ minimizing the objective.

You will use these models to predict the MPG (miles per gallon) of a car from several attributes of the vehicle. The data set is contained in the enclosed MATLAB file *autompg.mat* [1]. You can load the data in MATLAB by typing "load('autompg.mat');": this command will load four variables in the workspace: *trainsetX*, *trainsetY*, *testsetX*, and *testsetY*. Each row in these matrices corresponds to one example: *trainsetX(i,:)* contains the input vector of training example $i$, and *trainsetY(i)* the corresponding output value. Look at *autompg.names* for a description of the features. You should train your algorithm on the training set (*trainsetX*, *trainsetY*), and evaluate its performance on the test set (*testsetX*, *testsetY*). For all experiments, you should measure the performance as mean squared error on the test set, i.e., evaluate it by computing $\frac{1}{M}\sum_{i=1}^{M}\left[\hat{y}^{(i)} - f_\theta(\hat{x}^{(i)})\right]^2$ where $(\hat{x}^{(i)}, \hat{y}^{(i)})_{i=1,..,M}$ are the $M$ test examples, and $f_\theta$ is the function learned on the training set.

(a) [14 points] The performance of the algorithm will depend on the choice of the parameter $\lambda$. Tune this parameter via 10-fold cross validation using $\lambda \in \{10^{-5}, 10^{-3}, 10^{-1}, 10, 10^3, 10^5, 10^7\}$ for both the linear model and the quadratic model. Plot the cross-validation scores (computed as the average of the mean squared errors obtained over the 10 validation sets) as a function of $\lambda$ for both models.

(b) [4 points] Are there any values of $\lambda$ producing underfitting? If yes, which values?

(c) [4 points] Are there any values of $\lambda$ producing overfitting? If yes, which values?

[1] These examples were derived from the Auto-Mpg data set available at the UC Irvine Machine Learning Data Repository.

(d) [4 points] Which of the two versions of feature vector $b(x)$ produces more overfitting, $b^l(x)$ or $b^q(x)$? Can you explain why?

(e) [4 points] Now, plot the *test set* mean squared error for the same set of values of $\lambda$. Is the cross-validation score a good predictor of performance on the test set?

6. *Problem for graduate credit only: if you are enrolled in COSC74 (instead of COSC174) you do not need to solve this problem.*

**[25 points]** For this problem you need to implement locally weighted linear regression (LWLR) and to apply it to the MPG regression problem above. LWLR predicts the output of test example $x$ as $f_\theta(x) = \theta(x)^T b^l(x)$, where $\theta(x)$ is given by:

$$\theta(x) = \arg\min_\theta E^{LWLR}(\theta) \tag{3}$$

with $E^{LWLR}(\theta) = \frac{1}{2}\sum_{i=1}^m w^i(x)\left[y^{(i)} - \theta^T b^l(x^{(i)})\right]^2$ and $w^i(x) = \exp\left(-\frac{||x^{(i)}-x||^2}{2\tau^2}\right)$.

(a) [9 points] Show that the locally weighted linear regression error $E^{LWLR}(\theta)$ can be written in matrix notation as

$$E^{LWLR}(\theta) = (B\theta - \mathbf{y})^T W (B\theta - \mathbf{y}) \tag{4}$$

where $B$ and $\mathbf{y}$ are defined as in the slides used in class, and $W$ is a matrix whose entries are written as functions of the weights $w^i(x)$. Specify in detail what $W$ is.

(b) [10 points] It is easy to show (you may want to do it as an exercise) that the minimizer $\theta(x)$ of Eq. 4 can be computed via the following closed-form solution:

$$\theta(x) = (B^T W B)^{-1} B^T W \mathbf{y} \tag{5}$$

Use this closed-form solution to implement LWLR. In order to avoid numerical problems, before computing $W$, scale the weights $w^i(x)$ so that they sum up to 1 for each $x$. Plot the resulting 10-fold cross-validation score as a function of $\tau$ for $\tau \in \{10^2, 10^3, 10^5, 10^6\}$. Note that in each validation run, you will need to evaluate $M$ distinct functions $f_{\theta(\hat{x}^{(i)})}$, as the function itself varies with the test input $\hat{x}^{(i)}$.

(c) [2 points] Identify the values of $\tau$ (if any) causing underfitting.

(d) [2 points] Which values of $\tau$ yield overfitting?

(e) [2 points] Plot the test set error as a function of $\tau$. Is the performance similar to the one measured via cross validation?