Winter 2017: EECS 545 Homework 1

Due: 10 February 2017, 11:59 PM EST

Homework Policy: Working in groups is fine, but each member must submit their own writeup. Please write the members of your group on your solutions. There is no strict limit to the size of the group but we may find it a bit suspicious if there are more than 4 to a team. Questions labelled with (Challenge) are not strictly required, but you'll get some participation credit if you have something interesting to add, even if it's only a partial answer. For coding problems, please report your results (values, plots, etc.) in your written solution, and append the code in the end.

1 Probability

(10 points)

(1) (5 points) For the following equations, describe the relationship between them. Write one of four answers: "=", "$\geq$", "$\leq$" or "depends" to replace "?" in the following relations. Choose the most specific one and briefly explain why. Assume all probabilities are non-zero.

i $P(H = h|D = d)$? $P(H = h)$

ii $P(H = h|D = d)$? $P(D = d|H = h)P(H = h)$

(2) (5 points) Random variables X and Y have a joint distribution $p(x, y)$ and assume all the distributions are continuous. Prove the following:

i $\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}_X[X|Y]]$

ii $var[X] = \mathbb{E}_Y[var_X[X|Y]] + var_Y[\mathbb{E}_X[X|Y]]$

2 Maximum Likelihood Estimation

(15 points) Consider a random variable $\mathbf{X}$ (possibly a vector) whose distribution (probability density function or probability mass function) belongs to a parametric family. The density or mass function may be written as $f(\mathbf{x}; \theta)$, where θ is called the parameter, and can be either a scalar or vector. For example, in the univariate Gaussian distribution, θ can be a two dimensional vector consisting of the mean and the variance. Suppose the parametric family is known, but the value of the parameter is unknown. It is often of interest to estimate this parameter from observations of $\mathbf{X}$.

Maximum likelihood estimation is one of the most important parameter estimation techniques. Let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be i.i.d. (independent and identically distributed) random variables distributed according to $f(\mathbf{x}; \theta)$. By independence, the joint distribution of the observations is the product

$$L(\theta) = \prod_{i=1}^{n} f(\mathbf{X}_i; \theta) \ . \tag{1}$$

Viewed as a function of θ, this quantity is called the *likelihood* of θ. It is often more convenient to work with the *log-likelihood*,

$$\ell(\theta) = \sum_{i=1}^{n} \log f(\mathbf{X}_i; \theta) \ . \tag{2}$$

A maximum likelihood estimate (MLE) of θ is any parameter

$$\hat{\theta} \in \operatorname*{argmax}_{\theta} \sum_{i=1}^{n} \log f(\mathbf{X}_i; \theta) \ , \tag{3}$$

where "argmax" denotes the set of all values achieving the maximum. If the maximizer is unique, $\hat{\theta}$ is called the maximum likelihood estimate of θ.

(1) (10 points) Consider n i.i.d. 1-dimensional Gaussian random variables $X_1, \cdots, X_n$ each with mean α and variance σ^2. That is,

$$f(x; \alpha, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\alpha)^2}{2\sigma^2}\right) \ . \tag{4}$$

Find the maximum likelihood estimates of α and σ^2.

(2) (5 points) Let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be i.i.d. d-dimensional Gaussian random variables distributed according to $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. That is,

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d \, |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \ . \tag{5}$$

Find the maximum likelihood estimate of the mean vector $\boldsymbol{\mu}$. (You can assume the covariance matrix $\boldsymbol{\Sigma}$ is known.)

3 Weighted Least Squares Regression

(15 points)

(1) (10 points) Consider the following scenario: you manage to obtain a data set $\{(\mathbf{x}_i, y_i)_{i=1}^n\}$, where $\mathbf{x}_i \in \mathbb{R}^m$, which is to say that you have multivariate input vectors $\mathbf{x}_i$ and scalar outputs y_i. Model this as a Linear Regression (with offset) problem with parameter vector β and offset b. consider weights $\{c_i\}_{i=1}^n$, such that $c_i \geq 0 \ \forall i$. Find a solution for the following problem:

$$\underset{\mathbf{w}, b}{\text{minimize}} \ \|c_i(y_i - \beta^T \mathbf{x}_i - b)\|_2^2 \tag{6}$$

Argue that, if $c_i = 1 \ \forall i$, then this is equivalent to finding the maximum-likelihood estimate of $\mathbf{w} = [b \ \ \beta]^T$, with y_i being modeled as

$$y_i = \beta^T \mathbf{x}_i + b + \epsilon_i \ \forall i \tag{7}$$

where $\epsilon_i | \mathbf{x}_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. random variables and $\mathbf{x}_i$ has been observed for each i.

Hint: Arrange the c_i values into a matrix.

(2) (5 points) What would happen if $\epsilon_i | \mathbf{x}_i \sim \mathcal{N}(0, \sigma_i^2)$, which means that the noise has different variance for each i?

4 Bias-Variance Decomposition

(10 points) The ultimate goal of machine learning is to generalize, i.e. to achieve low expected error on unseen data. How can we measure error on data that are yet unseen? By decomposing the expected error into bias, variance, and noise, we can obtain insights into the ingredients of the expected error, which can inform us on how to control it.

In this problem we investigate the expected squared error of a regression function. Suppose we draw i.i.d training data set $D = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ from a stochastic data generation process $y = f(x) + \epsilon$, where f is an unknown function and ϵ is an additive random noise with zero mean and finite variance. Our goal is to learn a regression function $f_D(x)$ from D. The subscript D emphasizes that a different training set D will likely train a different regression function. The expected squared error of f_D is

$$\mathbb{E}_{XY,D}\left[\left(y - f_D(x)\right)^2\right] \tag{8}$$

where the expectation is taken with respect to two independent sources of randomness: the noise in data generation process, as well as the random draw of a particular training set D (a random subset of the entire population).

(1) (3 points) For any given x, show that

$$\mathbb{E}_{Y|x,D}\left[(y - f_D(x))^2\right] = \mathbb{E}_{Y|x}\left[(y - f(x))^2\right] + \mathbb{E}_D\left[(f(x) - f_D(x))^2\right] \ . \tag{9}$$

(2) (1 point) Show that

$$\mathbb{E}_{Y|x}\left[(y - f(x))^2\right] = var\left[\epsilon\right] \ . \tag{10}$$

We call $var\left[\epsilon\right]$ the *noise*, which is inherent in the data generation process and we cannot control.

(3) (3 points) For any given x, show that

$$\mathbb{E}_D\left[(f(x) - f_D(x))^2\right] = \left(\mathbb{E}_D\left[f_D(x)\right] - f(x)\right)^2 + \mathbb{E}_D\left[(f_D(x) - \mathbb{E}_D\left[f_D(x)\right])^2\right] \tag{11}$$

We define $bias(x) := \mathbb{E}_D\left[f_D(x)\right] - f(x)$. It measures in expectation, how far the predicted value deviates from the true expected value, due to our assumption on the functional form of f_D. The more flexible f_D is, the more likely f_D will match f, thus the smaller *bias*.

We define $variance(x) := \mathbb{E}_D\left[(f_D(x) - \mathbb{E}_D\left[f_D(x)\right])^2\right]$. It measures the variation of predicted value $f_D(x)$ caused by the random choice of training set D. The larger D is, the less random f_D is (in the limit, an infinitely large D is the entire population, which has no randomness), thus the smaller *variance*.

(4) (3 points) Now, show the bias-variance decomposition:

$$\mathbb{E}_{XY,D}\left[(y - f_D(x))^2\right] = var[\epsilon] + \mathbb{E}_X\left[(bias(x))^2\right] + \mathbb{E}_X\left[variance(x)\right] \ . \tag{12}$$

5 Linear Regression

(20 points) The files `xTrain.dat`, `yTrain.dat`, `xTest.dat`, and `yTest.dat` specify a linear regression problem. `xTrain.dat` represents the inputs ($\mathbf{x}_i \in \mathbb{R}$) and `yTrain.dat` represents the outputs ($y_i \in \mathbb{R}$) of the training set, with one training example per row.

(1) (5 points) Use the closed-form solution for linear regression (with offset) derived in class to perform Linear Regression on this training data. Use your favorite software package (Python, R, Matlab) with appropriate linear algebra functions for this problem. Please do NOT use any built-in or off-the-shelf functions for solving the Linear Regression problem. Give the coefficients generated by the closed-form solution.

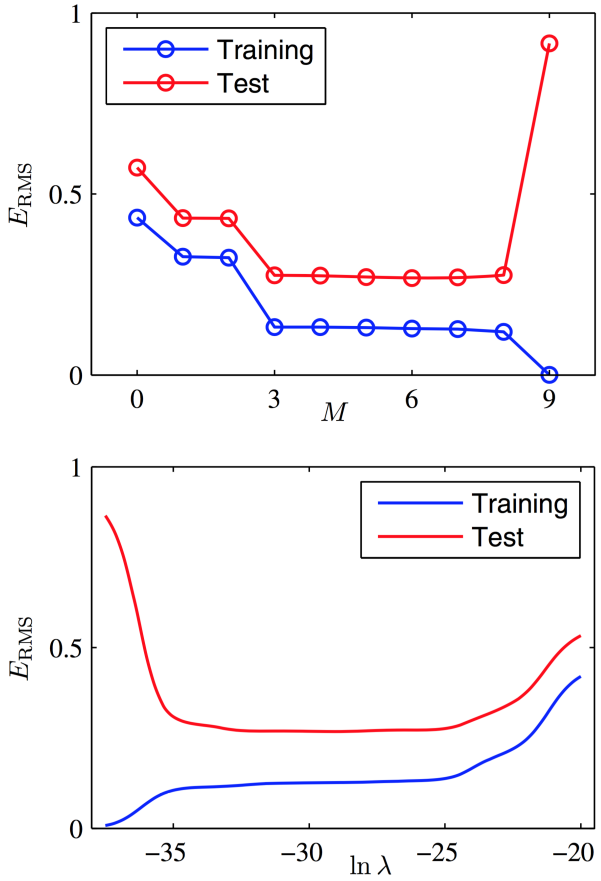
(2) (10 points) Next, you will investigate the problem of over-fitting. Recall the figure from the lecture that explored over-fitting as a function of the degree of polynomial M, where the root-mean-square (RMS) error is defined as

$$E_{RMS} = \sqrt{\frac{2E(\mathbf{w}^*)}{n}} \ , \tag{13}$$

where

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}\left(\sum_{j=0}^{M} w_j\phi_j(\mathbf{x}_i) - y_i\right)^2 = \frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{w}^\top\phi(\mathbf{x}_i) - y_i\right)^2 \ , \tag{14}$$

i Using the closed-form solution, find the coefficients for a M-degree polynomial (for $M = 0, \cdots, 9$) for the training data specified in `xTrain.dat` and `yTrain.dat`. Now use these parameters to generate the above chart, using `xTest.dat`, and `yTest.dat` as the test data.



ii Which degree polynomial would you say best fits the data? Was there evidence of under/over-fitting the data? Use your generated chart to defend your answer.

(3) (10 points) Finally, you will explore the role of regularization. Recall the figure from the lecture that explored the effect of the regularization weight λ:

i Again using the closed-form solution, find the coefficients for a 9th degree polynomial ($M = 9$) given regularization weight λ (for $\lambda = \{0, 10^{-6}, 10^{-5}, \cdots, 10^{-1}, 10^{0}(1)\}$) for the training data specified in `xTrain.dat` and `yTrain.dat`. Now use these parameters to generate the above chart, using `xTest.dat`, and `yTest.dat` as the test data. Specifically, use the following regularized objective function in finding the coefficient $\mathbf{w}$:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{w}^{\top}\phi(\mathbf{x}_i) - y_i\right)^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \ , \tag{15}$$

but remember to use the original E_{RMS} (Equation (13)-(14)) for plotting.

ii Which λ value seemed to work the best? Use your generated chart to defend your answer.