Winter 2017: EECS 545 Homework 1 Solutions

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

Depends. If $H = h$ and $D = d$ are independent events, then the two expressions are equal. If $H \cap D \neq \phi$, then either $P(H = h|D = d) \leq P(H = h)$ (for a small overlap between distributions D and H) or $P(H = h|D = d) \geq P(H = h)$ (if the event $D = d$ is a subset of $H = h$).

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

"$\geq$". Use Bayes' Rule to show that the left-hand side is equal to the right-hand side divided by $P(D = d)$, which here is $0 < P(D = d) \leq 1$.

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

$$\begin{aligned}\mathbb{E}[X] &= \int_Y \int_X x\ p(x, y)\ dx\ dy \\ &= \int_Y \left(\int_X x\ p(x|y)\ dx \right) p(y) dy \\ &= \mathbb{E}_Y[\mathbb{E}_X[X|Y]]\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}_Y[var_X[X|Y]] &= \mathbb{E}_Y[\mathbb{E}_X[X^2|Y]] - \mathbb{E}_Y[\mathbb{E}_X[X|Y]^2] \\ var_Y[\mathbb{E}_X[X|Y]] &= \mathbb{E}_Y[\mathbb{E}_X[X|Y]^2] - (\mathbb{E}_Y[\mathbb{E}_X[X|Y]])^2\end{aligned}$$

Add these and apply the result derived in $2(i)$ to obtain the required result.

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.

Solution: The log-likelihood function is

$$\ell(\mathbf{x}; \alpha, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(x_i - \alpha)^2}{2\sigma^2}$$

Differentiating this with respect to α and setting it equal to zero, we obtain

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

Differentiating this once again with respect to α, we obtain

$$\frac{\delta^2 \ell(\mathbf{x}; \alpha, \sigma^2)}{\delta \alpha^2} = -\frac{n}{\sigma^2} \leq 0$$

Which means that we have achieved the maximum. This is a global maximum because this condition holds everywhere.
Differentiating the log-likelihood function with respect to σ^2 and setting to zero, we obtain

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

Differentiating once more with respect to σ^2 we obtain

$$\frac{\delta^2 l(\mathbf{x}; \alpha, \sigma^2)}{\delta (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

So, our estimate works only for such σ^2 values as the ones which will cause this to be negative (i.e., when σ^2 is large). We have therefore reached a local maximum.

(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.)
Solution: We may apply an analysis similar to part (1) to obtain $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$.
Note that the Hessian of $\boldsymbol{\mu}$ is $\nabla^2_{\boldsymbol{\mu}} \ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\boldsymbol{\Sigma}^{-1}$. This matrix is negative definite since $\boldsymbol{\Sigma}$ is the covariance matrix of a multi-dimensional Gaussian distribution, which is positive definite. Therefore the log-likelihood function is concave and $\hat{\boldsymbol{\mu}}$ is the global maximum.

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:

$$\underset{\beta, b}{\text{minimize}} \ \sum_{i=1}^n c_i(y_i - \beta^T \mathbf{x}_i - b)^2 \tag{6}$$

Argue that, if $c_i = 1 \ \forall i$, then this is equivalent to finding the maximum-likelihood estimate of $\mathbf{w} = \begin{bmatrix} b \\ \beta \end{bmatrix}$, 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. Try to rewrite equation (7) in terms of $\mathbf{w}$.

Solution:

$$\begin{aligned} L(\beta, b) &= \sum_{i=1}^n c_i(y_i - \beta^T \mathbf{x}_i - b)^2 \\ &= \sum_{i=1}^n c_i(y_i - \mathbf{w}^T \tilde{\mathbf{x}}_i)^2 \end{aligned}$$

where $\tilde{\mathbf{x}}_i = \begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix}$.

Then, $L(\mathbf{w}) = \|C^{\frac{1}{2}}(\mathbf{y} - \tilde{X}\mathbf{w})\|_2^2$, with $C = diag(c_1, \ c_2, \ \cdots, \ c_n)$ and $\tilde{X} = \begin{bmatrix} 1 \ x_{11} & \cdots & x_{1m} \\ & \vdots & \\ 1 \ x_{n1} & \cdots & x_{nm} \end{bmatrix}$.

This is of the familiar linear regression form and so we may apply the same analysis and obtain $\mathbf{w} = (\tilde{X}^T C \tilde{X})^{-1} \tilde{X} C \mathbf{y}$.

We may find the joint probability density of the random vector $\mathbf{y}$ given $\tilde{X}$ to be

$$p(\mathbf{y}|\tilde{X}; \mathbf{w}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{\frac{-(y_i - \mathbf{w}^T \tilde{\mathbf{x}}_i)^2}{2\sigma^2}\right\}$$

Taking log, we notice that maximizing the log likelihood is the same as minimizing the loss function in (6).

(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?

Solution:

$$p(\mathbf{y}|\tilde{X}; \mathbf{w}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{\frac{-(y_i - \mathbf{w}^T \mathbf{x}_i)^2}{2\sigma_i^2}\right\} = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{\frac{-c_i(y_i - \mathbf{w}^T \mathbf{x}_i)^2}{2}\right\}$$

with $c_i = 1/\sigma_i^2$. So, this reduces to the weighted least squares regression problem.

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

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

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

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

Solution:

$$\begin{aligned}
&\mathbb{E}_{Y|x,D}\left[(y - f_D(x))^2\right] \\
&= \mathbb{E}_{Y|x,D}\left[(y - f(x) + f(x) - f_D(x))^2\right] \\
&= \underbrace{\mathbb{E}_{Y|x,D}\left[(y - f(x))^2\right]}_{(a)} + \underbrace{\mathbb{E}_{Y|x,D}\left[(f(x) - f_D(x))^2\right]}_{(b)} + \underbrace{2\mathbb{E}_{Y|x,D}\left[(y - f(x))(f(x) - f_D(x))\right]}_{(c)}
\end{aligned}$$

(a): Because $(y - f(x))$ is independent of D, $\mathbb{E}_{Y|x,D}\left[(y - f(x))^2\right] = \mathbb{E}_{Y|x}\left[(y - f(x))^2\right]$;

(b): Because $(f(x) - f_D(x))$ is independent of $Y|x$, $\mathbb{E}_{Y|x,D}\left[(f(x) - f_D(x))^2\right] = \mathbb{E}_D\left[(f(x) - f_D(x))^2\right]$;

(c): For independent random variables X and Y, $\mathbb{E}_{XY}[f(X)g(Y)] = \mathbb{E}_X[f(X)] \cdot \mathbb{E}_Y[g(Y)]$. Using the independence arguments in (a) and (b):

$$\begin{aligned}
&2\ \mathbb{E}_{Y|x,D}\left[(y - f(x))(f(x) - f_D(x))\right] \\
&= 2\ \mathbb{E}_{Y|x}\left[(y - f(x))\right] \cdot \mathbb{E}_D\left[(f(x) - f_D(x))\right] \\
&= 2\ \mathbb{E}_{Y|x}[\epsilon] \cdot \mathbb{E}_D\left[(f(x) - f_D(x))\right] \\
&= 2\ \cdot 0 \cdot \mathbb{E}_D\left[(f(x) - f_D(x))\right] \\
&= 0\ ,
\end{aligned} \tag{10}$$

where $\mathbb{E}_{Y|x}[\epsilon] = 0$ is defined in the problem setting. Therefore

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

for any given x.

(2) (1 point) Show that

$$\mathbb{E}_{Y|x}\left[(y - f(x))^2\right] = var[\epsilon] \ . \tag{11}$$

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

Solution:

$$\mathbb{E}_{Y|x}\left[(y - f(x))^2\right] = \mathbb{E}_{Y|x}\left[\epsilon^2\right] = \mathbb{E}_{Y|x}\left[(\epsilon - 0)^2\right] = \mathbb{E}_{Y|x}\left[(\epsilon - \mathbb{E}_{Y|x}[\epsilon])^2\right] = var[\epsilon] \ .$$

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

$$\mathbb{E}_D\left[(f(x) - f_D(x))^2\right] = (\mathbb{E}_D\left[f_D(x)\right] - f(x))^2 + \mathbb{E}_D\left[(f_D(x) - \mathbb{E}_D\left[f_D(x)\right])^2\right] \tag{12}$$

We define $\mathit{bias}(x) := \mathbb{E}_D\left[f_D(x)\right] - f(x)$. 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 $\mathit{variance}(x) := \mathbb{E}_D\left[(f_D(x) - \mathbb{E}_D\left[f_D(x)\right])^2\right]$. 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*.

Solution:

$$\begin{aligned}
&\mathbb{E}_D\left[(f(x) - f_D(x))^2\right] \\
=\ &\mathbb{E}_D\left[(f(x) - \mathbb{E}_D\left[f_D(x)\right] + \mathbb{E}_D\left[f_D(x)\right] - f_D(x))^2\right] \\
=\ &\underbrace{\mathbb{E}_D\left[(f(x) - \mathbb{E}_D\left[f_D(x)\right])^2\right]}_{(a)} + \underbrace{\mathbb{E}_D\left[(\mathbb{E}_D\left[f_D(x)\right] - f_D(x))^2\right]}_{(b)} + \underbrace{2\mathbb{E}_D\left[(f(x) - \mathbb{E}_D\left[f_D(x)\right])(\mathbb{E}_D\left[f_D(x)\right] - f_D(x))\right]}_{(c)}
\end{aligned}$$

(a): For a given x, both $\mathbb{E}_D\left[f_D(x)\right]$ and $f(x)$ are constants. Therefore $(f(x) - \mathbb{E}_D\left[f_D(x)\right])^2$ is also a constant. $\mathbb{E}_D\left[(f(x) - \mathbb{E}_D\left[f_D(x)\right])^2\right] = (f(x) - \mathbb{E}_D\left[f_D(x)\right])^2 = (\mathit{bias}(x))^2$.

(b): the term equals $\mathit{variance}(x)$;

(c): Using the same argument in (a), for a given x, $(f(x) - \mathbb{E}_D\left[f_D(x)\right])$ is a constant. Therefore

$$\begin{aligned}
&2\ \mathbb{E}_D\left[(f(x) - \mathbb{E}_D\left[f_D(x)\right])(\mathbb{E}_D\left[f_D(x)\right] - f_D(x))\right] \\
=\ &2\,(f(x) - \mathbb{E}_D\left[f_D(x)\right]) \cdot \mathbb{E}_D\left[\mathbb{E}_D\left[f_D(x)\right] - f_D(x)\right] \\
=\ &2\,(f(x) - \mathbb{E}_D\left[f_D(x)\right]) \cdot (\mathbb{E}_D\left[f_D(x)\right] - \mathbb{E}_D\left[f_D(x)\right]) \\
=\ &2\,(f(x) - \mathbb{E}_D\left[f_D(x)\right]) \cdot 0 \\
=\ &0\ .
\end{aligned}$$

Combining (a), (b), and (c) we have

$$\begin{aligned}
\mathbb{E}_D\left[(f(x) - f_D(x))^2\right] &= (\mathbb{E}_D\left[f_D(x)\right] - f(x))^2 + \mathbb{E}_D\left[(f_D(x) - \mathbb{E}_D\left[f_D(x)\right])^2\right] \\
&= (\mathit{bias}(x))^2 + \mathit{variance}(x)\ .
\end{aligned}$$

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

$$\mathbb{E}_{XY,D}\left[(y - f_D(x))^2\right] = \mathit{var}[\epsilon] + \mathbb{E}_X\left[(\mathit{bias}(x))^2\right] + \mathbb{E}_X\left[\mathit{variance}(x)\right]\ . \tag{13}$$

Solution: First we show that $\mathbb{E}_{XY}\left[g(x,y)\right] = \mathbb{E}_X\left[\mathbb{E}_{Y|x}\left[g(x,y)\right]\right]$. Suppose X, Y has joint probability density function $p(x,y)$, then following the definition of expectation, we have

$$\begin{aligned}
\mathbb{E}_{XY}\left[g(x,y)\right] &= \int_x \int_y g(x,y)p(x,y)dy\ dx \\
&= \int_x \int_y g(x,y)p(y|x)p(x)dy\ dx \\
&= \int_x \left(\int_y g(x,y)p(y|x)dy\right) p(x)dx \\
&= \int_x \mathbb{E}_{Y|x}\left[g(x,y)\right] p(x)dx \\
&= \mathbb{E}_X\left[\mathbb{E}_{Y|x}\left[g(x,y)\right]\right]\ .
\end{aligned}$$

Applying this "chain rule of expectation" to the left-hand side of (13), and combining the results so far in (1), (2), and (3), we have

$$\begin{aligned}
&\mathbb{E}_{XY,D}\left[(y - f_D(x))^2\right] \\
&= \mathbb{E}_X\left[\mathbb{E}_{Y|x,D}\,(y - f_D(x))^2\right] \\
&= \mathbb{E}_X\left[\mathbb{E}_{Y|x}\left[(y - f(x))^2\right] + (\mathbb{E}_D\left[f_D(x)\right] - f(x))^2 + \mathbb{E}_D\left[(f_D(x) - \mathbb{E}_D\left[f_D(x)\right])^2\right]\right] \\
&= \mathbb{E}_X\left[var[\epsilon] + (bias(x))^2 + variance(x)\right] \\
&= var[\epsilon] + \mathbb{E}_X\left[(bias(x))^2\right] + \mathbb{E}_X\left[variance(x)\right] \ .
\end{aligned}$$

The last line holds because the noise ϵ is independent of x as defined in the problem setting.

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.

Solution: (all solutions in this problem are up to numerical precisions of particular software packages.)

$y = 1.947 - 2.824x$.

(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

$$E_{RMS} = \sqrt{\frac{2E(\mathbf{w}^*)}{n}} \ , \tag{14}$$

where

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}\left(\sum_{j=0}^{M} w_j\phi_j(\mathbf{x}_i) - y_i\right)^2 = \frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{w}^\top\phi(\mathbf{x}_i) - y_i\right)^2 \ , \tag{15}$$

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.

Solution:

i The chart is shown in Figure 1.

ii 5-th degree polynomial best fits the data, since the regression function has lowest test error. Using 9-th degree polynomial *overfits* the data, since the training error is slow while the test error is very high. Using 0-2 degree polynomials *underfits* the data, since both training and test errors are high.

(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:

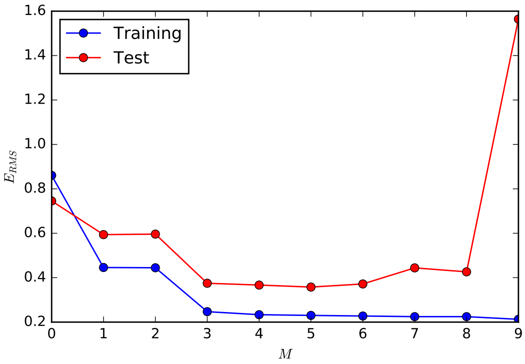

Figure 1: Training and test error with varying degree of polynomial.

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 $\mathbf{w}$:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{w}^\top\phi(\mathbf{x}_i) - y_i\right)^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \ , \tag{16}$$

but remember to use the original E_{RMS} (Equation (14)-(15)) for plotting.

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

Solution:

i The chart is shown in Figure 2.

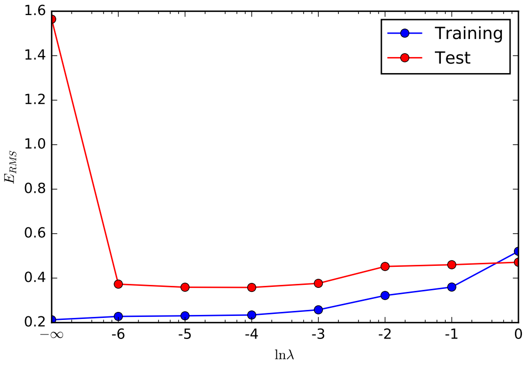

Figure 2: Training and test error with varying regularization coefficient.

ii $\lambda \in [10^{-5}, 10^{-4}]$ seems to be the “sweet spot” for this particular problem since it gives the lowest test error.