The University of Connecticut **Fall 2018**
Dept. of ECE **KRP**

<u>Take Home</u>
(Due December 10, 2018).

1. [10 points] Assume that x_n, n=1,2,…,N are i.i.d. observations from a Gaussian $N(\mu,\sigma^2)$. Obtain the MAP estimate of μ, if the prior follows the exponential distribution

$$p(\mu) = Exp(\mu;\lambda) = \lambda \exp(-\lambda\mu), \lambda > 0, \mu \geq 0$$

Obtain the Laplacian approximation of the posterior?

2. (10 points) Suppose we have features $\underline{x} \in R^p$, a two class response, with class sample sizes n_1, n_2 and the target responses $\{z_i\}$ coded as $-N/\ n_1$ for class 1, $N/\ n_2$ for class 2, where $N = n_1 + n_2$.

(a) Show that the linear discriminant analysis (LDA) rule classifies a test feature $\underline{x}$ to class 2 if

$$\underline{x}^T \hat{\Sigma}^{-1}(\underline{\hat{\mu}}_2 - \underline{\hat{\mu}}_1) > \frac{1}{2}\underline{\hat{\mu}}_2^T \hat{\Sigma}^{-1}\underline{\hat{\mu}}_2 - \frac{1}{2}\underline{\hat{\mu}}_1^T \hat{\Sigma}^{-1}\underline{\hat{\mu}}_1 + \ln\frac{n_1}{n_2}$$

and class 1 otherwise. Here

$$\hat{\mu}_i = \frac{1}{n_i}\sum_{k \in C_i} \underline{x}_k ; i = 1,2; C_i = \text{samples from class } i; |C_i| = n_i$$

$$\hat{\Sigma} = \frac{1}{N-2}\left(\sum_{i=1}^{2}\sum_{k \in C_i}(\underline{x}_k - \hat{\mu}_i)(\underline{x}_k - \hat{\mu}_i)^T\right)$$

(b) Consider minimization of the least squares criterion

$$J = \sum_{i=1}^{2}\sum_{k \in C_i}(z_i - w_0 - \underline{w}^T\underline{x})^2$$

Show that the solution $\hat{\underline{w}}$ satisfies

$$\left((N-2)\hat{\Sigma} + \frac{n_1 n_2}{N}\hat{\Sigma}_B\right)\underline{w} = N(\underline{\hat{\mu}}_2 - \underline{\hat{\mu}}_1)$$

where

$$\hat{\Sigma}_B = (\underline{\hat{\mu}}_2 - \underline{\hat{\mu}}_1)(\underline{\hat{\mu}}_2 - \underline{\hat{\mu}}_1)^T$$

(c) Show that

$$\hat{\underline{w}} \propto \hat{\Sigma}^{-1}(\underline{\hat{\mu}}_2 - \underline{\hat{\mu}}_1)$$

(d) Show that this result in (c) is valid for *any* distinct coding of the two classes.

(e) Find the solution $\hat{w}_0$ and hence the predicted responses $\hat{z}_i = \hat{w}_0 + \hat{\underline{w}}\,\underline{x}_i$.
Show that the decisions rule to classify to class 2 if $\hat{z}_i > 0$ and class 1 otherwise is not optimal unless the classes have equal number of observations.

3. (10 points) Let $z \sim N(z;\mu,\sigma^2)$. Show that

$$\mathrm{E}\left[\, z \mid z \geq c \right] = \mu + \sigma H\left(\frac{c-\mu}{\sigma}\right)$$

$$E\left[\, z^2 \mid z \geq c \right] = \mu^2 + \sigma^2 + \sigma(c+\mu) H\left(\frac{c-\mu}{\sigma}\right)$$

where

$$H(u) = \frac{\phi(u)}{1-\Phi(u)}$$

and where $\phi(u)$ is the pdf of a standard Gaussian and $\Phi(u)$ is its CDF.

4. [10 points] In this problem, you will prove that LMS converges in a mean square sense. Consider the LMS equation:

$$\underline{w}^{(n+1)} = \underline{w}^{(n)} + \eta(z^n - \underline{w}^{(n)T}\underline{x}^n)\underline{x}^n = \underline{w}^{(n)} + \eta(\underbrace{z^n - \underline{w}^{*T}\underline{x}^n}_{e^{*n}} - (\underline{w}^{(n)} - \underline{w}^*)^T\underline{x}^n)\underline{x}^n$$

$$\underline{v}^{(n+1)} == \left[I - \eta\underline{x}^n\underline{x}^{nT}\right]\underline{v}^{(n)} + \eta e^{*n}\underline{x}^n ; \underline{v}^{(n)} = \underline{w}^{(n)} - \underline{w}^*$$

(a) Let $\Sigma_n = E\{\underline{v}^{(n)}\underline{v}^{(n)T}\}; R_x = E[\underline{x}^n\ \underline{x}^{nT}] \sim$ Correlation matrix of data; $E[\left(e^{*n}\right)^2] = \sigma_e^2$

Using LMS assumption and the orthogonality of error and the weight estimate, show that

$$\begin{aligned}\Sigma_{n+1} &= \Sigma_n - \eta R_x\Sigma_n - \eta\Sigma_n R_x + \eta^2 E\{\underline{x}^n\underline{x}^{nT}\Sigma_n\underline{x}^n\underline{x}^{nT}\} + \eta^2 E\{\left(e^{*n}\right)^2\underline{x}^n\underline{x}^{nT}\} \\ &= \Sigma_n - \eta R_x\Sigma_n - \eta\Sigma_n R_x + 2\eta^2 R_x\Sigma_n R_x + \eta^2 R_x tr\{\Sigma_n R_x\} + \eta^2\sigma_e^2 R_x\end{aligned}$$

(Hint: Use the fourth order moment equations of Gaussian random variables)

(b) Consider the Eigen decomposition of $R_x = Q\Lambda_x Q^T$ and let $\hat{\Sigma}_{n+1} = Q^T\Sigma_{n+1}Q$

Show that $\hat{\Sigma}_{n+1} = \hat{\Sigma}_n - \eta\Lambda_x\hat{\Sigma}_n - \eta\hat{\Sigma}_n\Lambda_x + 2\eta^2\Lambda_x\hat{\Sigma}_n\Lambda_x + \eta^2\Lambda_x tr\{\hat{\Sigma}_n\Lambda_x\} + \eta^2\sigma_e^2\Lambda_x$

(c) Now consider the diagonal elements of $\hat{\Sigma}_{n+1}$ and represent them as a vector $\underline{s}_{n+1}$

Show that

$$\underline{s}_{n+1} = (I_{p+1} - 2\eta\Lambda_x + 2\eta^2\Lambda_x^2 + \eta^2\underline{\lambda}\,\underline{\lambda}^T)\underline{s}_n + \eta^2\sigma_e^2\underline{\lambda}$$

$$where\ \underline{\lambda} = \begin{bmatrix}\lambda_1 & \lambda_2 & . & \lambda_{p+1}\end{bmatrix}^T$$

(d) Show that this system is stable if

$$0 < \eta < \frac{2}{\sum_{i=1}^{p+1}\lambda_i} = \frac{2}{tr(R_x)}$$

5. (10 points) Consider a general regularized least squares regression problem.

$$J = \frac{1}{N} \| \underline{z} - X\,\underline{w} \|_2^2 + \frac{\lambda}{N} \underline{w}^T \Gamma^T \Gamma \underline{w};\ \underline{z} \in R^N;\ X \in R^{N\mathbf{x}(p+1)}$$

$$\text{where } \underline{z} = X\,\underline{w} + \underline{v};\ v_n \sim N(0,\sigma^2)\ \forall n = 1,2,..,N$$

$$\text{Let } \hat{\underline{w}}(0,\Gamma) = (X^T X)^{-1} X^T \underline{z},\ \text{least squares solution when } \lambda = 0.$$

a) Show that the optimal solution is a biased estimate given by

$$\hat{\underline{w}}(\lambda,\Gamma) = \underline{w} - \lambda (X^T X + \lambda \Gamma^T \Gamma)^{-1} \Gamma^T \Gamma \underline{w} + (X^T X + \lambda \Gamma^T \Gamma)^{-1} X^T \underline{v}$$

Specialize the estimate when $\Gamma = I_{p+1}$ and $\Gamma = X$. The latter is called uniform weight decay. Why? (Hint: It is related to $\hat{\underline{w}}(0,\Gamma)$.)

b) Show that the bias in the weight estimate is given by

$$\underline{w} - E_{\underline{v}}\{\hat{\underline{w}}(\lambda,\Gamma)\} = \lambda (X^T X + \lambda \Gamma^T \Gamma)^{-1} \Gamma^T \Gamma \underline{w}$$

Specialize the expected bias estimate when $\Gamma = I_{p+1}$ and $\Gamma = X$. Show that the bias is only a function of λ and $\underline{w}$ when $\Gamma = X$.

c) Show that the residual for a test vector ($\underline{x}$, z) is given by

$$r = z - \hat{z} = \underline{x}^T \underline{w} + v - \underline{x}^T \hat{\underline{w}}(\lambda,\Gamma) = \lambda \underline{x}^T (X^T X + \lambda \Gamma^T \Gamma)^{-1} \Gamma^T \Gamma \underline{w} + v - \underline{x}^T (X^T X + \lambda \Gamma^T \Gamma)^{-1} X^T \underline{v}$$

Specialize the residual expression for $\Gamma = I_{p+1}$ and $\Gamma = X$.

d) Now, we compute square of the bias of the residual assuming the second moment matrix $\Sigma_x = E_{\underline{x}}(\underline{x}\underline{x}^T) \approx \frac{X^T X}{N}$. Show that

$$bias^2(\lambda,\Gamma) = E(r)^2 \approx \lambda^2 \underline{w}^T \Gamma^T \Gamma (N\Sigma_x + \lambda \Gamma^T \Gamma)^{-1} \Sigma_x (N\Sigma_x + \lambda \Gamma^T \Gamma)^{-1} \Gamma^T \Gamma \underline{w}$$

When $\Gamma = I_{p+1}$ and $\Sigma_x = I_{p+1}$, show that

$$bias^2(\lambda, I_{p+1}) \approx \frac{\lambda^2}{(\lambda + N)^2} \underline{w}^T \underline{w}$$

Further when $\Gamma = X$ and $\Sigma_x = I_{p+1}$, show that

$$bias^2(\lambda, X) \approx \frac{\lambda^2}{(\lambda + 1)^2} \underline{w}^T \underline{w}$$

e) Show that, under the same assumption as in (d), the variance of the residuals is given by

$$\mathrm{var}(\lambda,\Gamma) = E\{[r - E(r)]^2\} = \sigma^2 + [E_{\underline{x},\underline{v}}\{\underline{x}^T (X^T X + \lambda \Gamma^T \Gamma)^{-1} X^T \underline{v}\underline{v}^T X (X^T X + \lambda \Gamma^T \Gamma)^{-1} \underline{x}]$$

$$\approx \sigma^2 (1 + N.tr([\Sigma_x (N\Sigma_x + \lambda \Gamma \Gamma^T)^{-1}]^2)$$

When $\Gamma = I_{p+1}$ and $\Sigma_x = I_{p+1}$, show that

$$\mathrm{var}(\lambda, I_{p+1}) \approx \sigma^2[1 + \frac{(p+1)N}{(N+\lambda)^2}]$$

Further when $\Gamma = X$ and $\Sigma_x = I_{p+1}$, show that

$$\mathrm{var}(\lambda, X) \approx \sigma^2[1 + \frac{(p+1)}{N(1+\lambda)^2}]$$

f) Find the optimal λ that minimizes the mean square error = (bias^2 +variance) for the two cases: (i) $\Gamma = I_{p+1}$ and $\Sigma_x = I_{p+1}$ and (ii) $\Gamma = X$ and $\Sigma_x = I_{p+1}$.

6. (10 points) (a) Consider a support vector machine and the following training data from two categories:

$$C_1 : \left\{ \underline{x}^1 = \begin{bmatrix} 1 \\ 5 \end{bmatrix}; \underline{x}^2 = \begin{bmatrix} -2 \\ -4 \end{bmatrix} \right\}$$

$$C_2 : \left\{ \underline{x}^3 = \begin{bmatrix} 2 \\ 3 \end{bmatrix}; \underline{x}^4 = \begin{bmatrix} -1 \\ 5 \end{bmatrix} \right\}$$

(i) Use the map $\underline{\Phi}(\underline{x})$ to map $\underline{x}$ to a higher dimensional space

$$\underline{\Phi}(\underline{x}) = [1 \sqrt{2}x_1 \ \sqrt{2}x_2 \ \sqrt{2}x_1x_2 \ x_1^2 \ x_2^2 \ ]^T$$

(ii) Formulate the dual problem associated with the SVM classification problem and solve it by hand. Check your answers with MATLAB or any SVM tool box you may have access to.

(iii) Find the discriminant function $g(x_1,x_2) = 0$ in the x_1-x_2 plane. Identify the support vectors from $g(x_1,x_2) = \pm 1$.

(iv) What is the margin? (Hint: Use result from Problems 7.4 and 7.5 of Bishop).

7. (10 points) Consider fitting a model of the form

$$p(z \mid x) = N(z; w_0 + w_1 x, \sigma^2)$$

Suppose we have made N= 11 measurements given by

$\{x^n\}$ = [94, 96, 94, 95, 104, 106,108, 113, 115, 121, 131]
$\{z^n\}$ = [0.47,0.75,0.83,0.98,1.18,1.29,1.40,1.60,1.75,1.90,2.23]

(a) Compute an unbiased estimate of σ^2 based on MLE estimate of $\underline{\hat{w}} = \begin{bmatrix} \hat{w}_0 \\ \hat{w}_1 \end{bmatrix}$

(b) Suppose the prior on $\underline{w}$ is of the form

$$p(\underline{w}) = N(\underline{0}, Diag(10^{10}, 1)).$$

Compute the marginal posterior of the slope $p(w_1 \mid \{x^n\}, \{z^n\}, \hat{\sigma}^2\}$. Here, $\hat{\sigma}^2$ is the estimate of variance from (a). Compute the mean and variance of the marginal posterior of the slope.

8. (10 points) Consider the negative log of the posterior given by

$$J = -\ln p(\theta_1, \theta_2 \mid D) = N\theta_2 + \frac{e^{-2\theta_2}}{2}\left[Ns^2 + N(\bar{z} - \theta_1)^2 \right]$$

where $\bar{z}$ is the sample mean and s^2 is the sample variance.

(a) Compute the gradient and Hessian of J and compute the MAP estimates of the parameters.

(b) Use this to derive a Laplace approximation of the posterior $p(\theta_1, \theta_2 \mid D)$.

9. [10 points] Consider a cause-effect model where the set of binary variables $\{h_1, h_2, \ldots, h_m\}$ are the causes (hidden or latent variables) and the set of binary variables $\{v_1, v_2, \ldots, v_n\}$ are the effects (visible or observed variables) with the joint distribution given by

$$P(\underline{v}, \underline{h}) = \frac{1}{Z} \exp\left(\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} h_i v_j + \sum_{i=1}^{m} b_i h_i + \sum_{j=1}^{n} c_j v_j\right)$$

$$where\ Z = \sum_{\underline{v}}\sum_{\underline{h}} \exp\left(\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} h_i v_j + \sum_{i=1}^{m} b_i h_i + \sum_{j=1}^{n} c_j v_j\right)$$

(a) Show that $P(\underline{h}|\underline{v})$ is given by

$$P(\underline{h} \mid \underline{v}) = \prod_{i=1}^{m} \frac{\exp\left(\sum_{j=1}^{n} d_{ij} v_j + b_i\right) h_i}{\left[1 + \exp\left(\sum_{j=1}^{n} d_{ij} v_j + b_i\right)\right]}; h_i \in \{0,1\}$$

and consequently

$$P(h_i = 1 \mid \underline{v}) = \frac{\exp\left(\sum_{j=1}^{n} d_{ij} v_j + b_i\right)}{\left[1 + \exp\left(\sum_{j=1}^{n} d_{ij} v_j + b_i\right)\right]} = g\left(\sum_{j=1}^{n} d_{ij} v_j + b_i\right) \ldots sigmoid\ \ function$$

(b) By symmetry, show that

$$P(\underline{v} \mid \underline{h}) = \prod_{j=1}^{n} \frac{\exp\left(\sum_{i=1}^{m} d_{ij} h_i + c_j\right) v_j}{\left[1 + \exp\left(\sum_{i=1}^{m} d_{ij} h_i + c_j\right)\right]}; v_j \in \{0,1\}$$

and consequently

$$P(v_j = 1 \mid \underline{h}) = \frac{\exp\left(\sum_{i=1}^{m} d_{ij} h_i + c_j\right)}{\left[1 + \exp\left(\sum_{i=1}^{m} d_{ij} h_i + c_j\right)\right]} = g\left(\sum_{i=1}^{m} d_{ij} h_i + c_j\right)$$

10. [10 points] Consider the problem of clustering one dimensional data with a mixture of 2 Gaussians using the EM algorithm. You are given three points $x^1=1$, $x^2=10$, $x^3=20$. Suppose that the output of the E-step is the following responsibility matrix:

$$\Gamma = \begin{bmatrix} 1 & 0 \\ 0.4 & 0.6 \\ 0 & 1 \end{bmatrix}$$

where the entry γ_{nk} is the probability of observation x^n belongs to cluster k. You are asked to compute the M-step.

(a) Write down the likelihood function you are trying to optimize.
(b) Perform the M-step for the two mixing weights and the two means.
(c) Find the final converged mixing weights, means and responsibilities.