The University of Connecticut **Spring 2021**
Dept. of ECE **KRP**

<u>Take Home</u>
(Due April 27, 2021).

1. (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

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

   and class 1 otherwise. Here

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

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

   (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

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

   where

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

   (c) Show that

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

   (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.

2. (10 points) Consider the following observation model where $\underline{z}$ is an unknown latent vector of dimension m, $\underline{x}$ is a measurement vector of dimension n, H is the unknown measurement matrix and the noise variance σ^2 is unknown as well.

$$\underline{x} = H\underline{z} + \underline{v};\ \underline{z} = N(\underline{0}, I_m);\ \underline{v} = N(\underline{0}, \sigma^2 I_n)$$

We are given N measurements $D^N = \{\underline{x}_n\}_{n=1}^N$ and the following two methods are suggested for solving this problem.

a) In the maximum likelihood method, one evaluates the density of $\underline{x}$ and optimizes the likelihood function $p(D^N \mid A, \sigma^2)$. Find the optimal A_{ML} and σ^2_{ML}.

b) Another way is to use the EM algorithm by forming the complete log-likelihood function $\ln p(D^N, Z^N \mid A, \sigma^2)$ where $Z^N = \{\underline{z}_n\}_{n=1}^N$. Here, we use the E-step to estimate the conditional mean $E(\underline{z}_n \mid \underline{x}_n)$ and conditional covariance $\Sigma_{z|x}$ and the M-step to update A and σ^2. Derive the relevant equations for the E and M steps.

3. [15 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

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

4. (15 points) Consider a general regularized least squares regression problem.

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

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

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

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

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

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

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

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

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

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

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 $\Sigma_x = E_{\underline{x}}(\underline{x}\underline{x}^T) \approx \frac{X^TX}{N}$. 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

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

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

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 $\Sigma_x = I_{p+1}$.

5. (15 points)
   a. Show that the value of the margin M for the maximum margin hyperplane in SVM is given by the following three relations:

   $$\frac{1}{M^2} = \underline{w}^{*T} \underline{w} = 2q(\underline{\lambda}) = \sum_{n=1}^{N} \lambda_n$$

   where $q(\underline{\lambda})$ is the dual function associated with the Lagrangian function of SVM classifier

   $$L(\underline{w}, w_0, \underline{\lambda}) = \frac{1}{2} \underline{w}^T \underline{w} - \sum_{n=1}^{N} \lambda_n \{ z^n (\underline{w}^T \underline{\phi}(\underline{x}^n) - w_0) - 1 \}$$

   b. 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

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

   (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?

6. (5 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 $p(\theta_1,\theta_2 \mid D)$.

7. [10 points] Let $x \in \{0,1\}$ denote the state of a component with probability $P(x=1) = \theta_1$. Suppose an imperfect test is performed and you get to observe its outcome, $y \in \{0,1\}$. The result is correct with probability $P(y \mid x, \theta_2)$ is given by the contingency table (binary symmetric channel model):

$x \backslash y$	$y = 0$	$y = 1$
$x = 0$	θ_2	$1-\theta_2$
$x = 1$	$1-\theta_2$	θ_2

a. Write down the joint probability mass function $P(x,y|\theta_1, \theta_2)$ as a 2x2 contingency table.
b. Suppose we have the following dataset: $\underline{x} = \{1,1,0,1,1,0,0\}$ and $\underline{y} = \{1,0,0,0,1,0,1\}$. What are the ML estimates of θ_1, θ_2? What is $p(D \mid \hat{\theta}_{1ML}, \hat{\theta}_{2ML}, M_2)$ where M_2 denotes this 2-parameter model?
c. Now consider a model with 4 parameters, $\underline{\theta} = (\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11})$ representing $P(x,y|\underline{\theta}) = \theta_{xy}$. What is the ML estimate of $\underline{\theta}$? What is $p(D \mid \hat{\underline{\theta}}_{ML}, M_4)$ where M_4 denotes the 4-paremeter model?
d. Suppose we are not sure which model is correct. We compute the leave-on-out cross validated log-likelihood of the 2-paremeter and the 4-parameter model as follows:

$$L(M) = \sum_{i=1}^{n} \ln p(x_i, y_i \mid \hat{\underline{\theta}}_{ML}(D_{-i}), M_4)$$

Where $\hat{\underline{\theta}}_{ML}(D_{-i})$ denotes the ML estimate computed on D excluding i^{th} measurement set. Which model will LOOV pick and why?
e. An alternative to LOOCV is to use the BIC criterion, defined as

$$BIC(M,D) = \ln P(D \mid \hat{\underline{\theta}}_{ML}, M) - \frac{dof(M)}{2} \ln N$$

Where $dof(M)$ is the number of free parameters in the model. Compute the BIC scores for both models. Which model does BIC prefer?

8. [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

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

and consequently

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

(b) By symmetry, show that

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

and consequently

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

9. [10 points] Consider a generalized mixture density where

$$p(\underline{x}_n\,|\,\underline{\theta}) = \sum_{j=1}^{m} p_j \sum_{k=1}^{l} q_k N(x_n\,|\,\mu_j, \sigma_k^2)$$

$$\textit{where } \underline{\theta} = \{p_1, p_2, \ldots, p_m, \mu_1, \mu_2, \ldots, \mu_m, q_1, q_2, \ldots, q_l, \sigma_1^2, \sigma_2^2, \ldots, \sigma_l^2\} \textit{ are all parameters.}$$

This density can be thought of having two latent variables $\underline{y}$ of dimension m and $\underline{z}$ of dimension l such that $p_j = P(y_n = j)$ and $q_k = P(z_n = k)$. We can think of this as a mixture of a mixture in the sense that for a given j, it is a mixture of Gaussian densities with different variances, but the same mean μ_j. EM algorithm is proposed for solving this problem.

a. Derive an expression for the responsibilities, $\pi_{njk} = P(y_n = j, z_n = k|x_n, \underline{\theta})$, $\gamma_{nj} = P(y_n = j|x_n, \underline{\theta})$ and $\delta_{nk} = P(z_n = k|x_n, \underline{\theta})$ needed for the E-step.

b. Write out the complete expression for the expected complete log-likelihood

$$Q(\underline{\theta}^{new}, \underline{\theta}^{old}) = E_{\underline{\theta}^{old}}\left\{\sum_{n=1}^{N} \ln P(y_n, z_n, x_n\,|\,\underline{\theta}^{new}\right\}$$

c. Solving the M-step would require us to jointly optimize the means and variances. Ths can be done in an iterative way by fixing the variances and solving for the means and vice versa. Derive the M-step.