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

Problem Set # 1

(Due September 17, 2018)

1. (15 points) (Calculus) Consider the following two nonlinear functions:

$$sigmoid\ \ function: \sigma(x) = \frac{1}{1+\exp(-x)}$$

$$softplus\ \ function: \varsigma(x) = \ln(1+\exp(x))$$

Softplus function is a continuously differentiable approximation to x^+ =max (0, x).

a) Show

$$(i)\ (1-\sigma(x)) = \sigma(-x)$$

$$(ii)\ \frac{d\sigma(x)}{dx} = \sigma(x)(1-\sigma(x)) = \sigma(x)\sigma(-x)$$

$$(iii)\ \ln\sigma(x) = -\varsigma(-x)$$

$$(iv)\ \frac{d\varsigma(x)}{dx} = \sigma(x)$$

$$(v)x = \ln\frac{\sigma(x)}{1-\sigma(x)}; \sigma(x) \in (0,1); x \in (-\infty, \infty)$$

$$(vi)x = \ln(\exp(\varsigma(x)) - 1); \varsigma(x) \in (0,\infty); x \in (-\infty, \infty)$$

$$(vii)\varsigma(x) - \varsigma(-x) = -\ln\sigma(-x) + \ln\sigma(x) = \ln\frac{\sigma(x)}{\sigma(-x)} = x$$

b) Compute the gradient and Hessian of the following functions with respect to $\underline{w}$. Here z_n and $\underline{x}_n$ are known for n=1, 2… N.

$$i.\ f(\underline{w}) = \sum_{n=1}^{N}(z_n - \underline{w}^T\underline{x}_n)^2; z_n \in R$$

$$ii.\ f(\underline{w}) = \sum_{n=1}^{N}(z_n - y_n)^2 \ \ where\ y_n = g(\underline{w}^T\underline{x}_n) = \frac{1}{1+e^{-\underline{w}^T\underline{x}_n}}; z_n \in \{0,1\}$$

$$iii.\ f(\underline{w}) = -\sum_{n=1}^{N}[z_n \log y_n + (1-z_n)\log(1-y_n)\ \ where\ y_n = g(\underline{w}^T\underline{x}_n) = \frac{1}{1+e^{-\underline{w}^T\underline{x}_n}}; z_n \in \{0,1\}$$

Is the function in (1) convex with respect to $\underline{w}$ (that is, Hessian is positive (semi) definite)? Is the function in (ii) convex? Is the function in (iii) convex? Check your answers for scalar $\{\underline{x}_n\}$.

2. (10 points) (ECE 6111 review: Bayes rule) These are three simple applications of Bayesian Inference.
   a) Using Bayes rule, prove the following logic statement: Given "If A is true then B is true", one may deduce that "if B is false, A is false".

b) Consider a "noisy" XOR gate with the conditional probabilities as shown in the following Table.

B	C	P(A=1|B,C)
0	0	0.10
0	1	0.99
1	0	0.80
1	1	0.25

Assume that events B and C are independent with prior probabilities P(B=1)=0.65 and P(C=1)=0.77. What is P(B=1|A=0), P(C=1|A=0), P(B=1|A=1) and P(C=1|A=1)?

c) Suppose there are two opaque bags, each containing 2 balls. It is known that one bag has 2 black balls and the other has a black ball and a white ball. You pick a bag at random and then one of the balls in that bag at random. When you look at the ball, it is black. You now pick the second ball from that same bag. What is the probability that this ball is also black? What is the probability that this ball is a white one?

d) Draw factor graphs for problems (b) and (c) and compute the requisite probabilities.

3. (15 points) (Review moments and functions of random variables) Let $X \sim$ Ga (a,b) where Ga denotes gamma density. That is,

$$Ga(x;a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}; x > 0; a > 0; b > 0$$

(1) *Compute* $E(x), Var(x), \text{mode}, E(\ln x)$ & *entropy* $H(x)$

(2) *Let* $Y = 1/X$. *Show that* $Y \sim IG(a,b)$, *i.e., inverse gamma density*

$$IG(y;a,b) = \frac{b^a}{\Gamma(a)} y^{-a-1} e^{-b/y}; y > 0; a > 0; b > 0$$

Wishart density is a generalization of gamma density to multiple dimensions, while inverse Wishart density is a generalization of inverse gamma density to multiple dimensions. Inverse gamma has a number of applications in survival analysis.

Hint: Exploiting the properties of exponential family of densities will make this and the next problem easier to solve.

For example, to get $E[\ln x]$, *note* $G(x;a,b) = e^{a\ln b - \ln\Gamma(a) + (a-1)\ln x - bx} = e^{\begin{bmatrix}(a-1) & -b\end{bmatrix}\begin{bmatrix}\ln x \\ x\end{bmatrix} - \underbrace{(\ln\Gamma(a) - a\ln b)}_{A(a,b)}}$

$$E(\ln x) = \frac{\partial A(a,b)}{\partial a} = \frac{d\ln\Gamma(a)}{da} - \ln b = \Psi(a) - \ln b;\ \Psi(a) = \frac{d\ln\Gamma(a)}{da} \ldots. \textit{digamma function}$$

$$Also, E(x) = -\frac{\partial A(a,b)}{\partial b} = \frac{a}{b}; \text{cov}(\ln x, x) = \begin{bmatrix} \frac{d^2\ln\Gamma(a)}{da^2} & \frac{1}{b} \\ \frac{1}{b} & \frac{a}{b^2} \end{bmatrix}$$

4. (10 points) (Review moments of random variables) Suppose θ ~ Beta (a,b), that is,

$$Beta(\theta;a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}; \theta \in [0,1]; a > 0; b > 0$$

Find E(θ), Var(θ), mode(θ) & entropy $H(\theta)$.

5. (10 points) (Kullback-Leibler divergence): Compute KL ($p \| q$) when $p(\underline{x})$ and $q(\underline{x})$ are multivariate Gaussian distributions? What happens when $p(\underline{x})$ is a weighted sum of Gaussian distributions? Discuss why it is difficult to compute when $q(\underline{x})$ is a weighted sum of Gaussian distributions. Do the problem first when x is a scalar and then generalize the scalar results to multivariate Gaussian distributions.
6. (5 points) (Information-theoretic Measures) Problem 1.39 of Bishop, Chapter 1, pp. 65.
7. (5 points) Using $I(X;Y|Z)=H(X|Z)-H(X|Y,Z)$, compute $I(X;Y|Z)$ when X,Y and Z are Gaussian random variables.
8. (5 points) Problem 2.43 of Bishop, Chapter 2, pp. 135.
9. (10 points) (Linear Regression) Consider a noisy target $z = \underline{w}^T \underline{x} + v$ for generating the data, where v is a noise term with zero mean and variance σ^2, *independently* generated for every sample $(\underline{x}, z)$.

   For the data $D = \{(\underline{x}_1, z_1), (\underline{x}_2, z_2) \ldots. (\underline{x}_N, z_N)\}$, denote the noise term in z_n as v_n and let $\underline{v} = [\, v_1, v_2 \ldots. v_N]^T$, $\underline{z} = [\, z_1, z_2, \ldots, z_N]^T$, $X = [\, \underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N]^T$ an N by p matrix. Assume that the p by p matrix $X^T X$ is invertible.
   Let the objective function to be minimized be

$$J(\underline{w}) = \frac{1}{N}\sum_{n=1}^{N}(z_n - \underline{w}^T \underline{x}_n)^2 = \frac{1}{N} \| \underline{z} - X\,\underline{w} \|_2^2$$

   a) Compute the optimal estimate $\hat{\underline{w}}$ that minimizes $J(\underline{w})$.

   b) Compute the optimal prediction $\hat{\underline{z}}$ and show that

$$\hat{\underline{z}} = X\,\hat{\underline{w}} = X\,\underline{w} + H\,\underline{v}; H = \underbrace{X\,(X^T X)^{-1} X^T}_{\text{Projection Matrix}}$$

   c) Show that the error $\underline{z} - \hat{\underline{z}} = (I_N - H)\underline{v}$ and that $trace(H)=p$.

   d) Show that

$$E[J(\underline{w})|_{\underline{w}=\hat{\underline{w}}}] = \frac{1}{N} E\{\| \underline{z} - \hat{\underline{z}} \|_2^2\} = \frac{1}{N} trace[(I_N - H)E\{\underline{v}\underline{v}^T\}(I_N - H)]$$

$$= \frac{\sigma^2}{N} trace(I_N - H) = \sigma^2[1 - \frac{trace(H)}{N}] = \sigma^2(1 - \frac{p}{N})$$

   e) Now suppose that we get test data $\underline{x}_{N+1}$ with a noisy target z_{N+1} and noise term v_{N+1}. Assume that the second moment matrix $\Sigma = E_{\underline{x}}[\underline{x}\underline{x}^T]$ is nonsingular. Show that the error

$$z_{N+1} - \hat{z}_{N+1} = z_{N+1} - \underline{x}_{N+1}^T \hat{\underline{w}} = (v_{N+1} - \underline{x}_{N+1}^T (X^T X)^{-1} X^T \underline{v})$$

   f) Show that

$$E[(z_{N+1} - \hat{z}_{N+1})^2] = E[(v_{N+1} - \underline{x}_{N+1}^T (X^T X)^{-1} X^T \underline{v})^2]$$

$$= \sigma^2 + \frac{\sigma^2}{N} trace[(\frac{1}{N} X^T X)^{-1} \Sigma] \approx \sigma^2(1 + \frac{p}{N})$$

10. (10 points) (A Simple Perceptron) Consider a two-dimensional plane. Choose a random line in the x_1-x_2 plane $w_0 + w_1 x_1 + w_2 x_2 = 0 \Rightarrow [\, w_0 \quad w_1 \quad w_2] \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} = \underline{w}^T \underline{x} = 0$ as your target function, where one side of the line sign $(\underline{w}^T \underline{x}) = 1$ maps to $z = +1$ (label *) and the other side sign $(\underline{w}^T \underline{x}) = -1$ maps to $z = -1$ (label o).

a) Generate a dataset $\{\underline{x}_n , z_n: n=1,2..,20\}$, that is, $N=20$. Plot the samples as well as the target function in the x_1-x_2 plane. Be sure to mark the examples from different classes differently, and add labels to the axes of the plot.

b) You want to learn the weights to classify the dataset correctly. Start with any $\underline{w}(0)$ at iteration $t=0$. At iteration t, the algorithm picks a sample from $\{\underline{x}_n , z_n: n=1,2..,20\}$ that is currently misclassified, call it $\{\underline{x}(t), z(t)\}$ and use it to update $\underline{w}(t)$. Since the sample is misclassified, we have $z(t) \neq \text{sign}(\underline{w}^T(t)\underline{x}(t))$, the update rule is a type of reinforcement learning ("training with a critic") of the form

$$\underline{w}(t+1) = \underline{w}(t) + z(t)\underline{x}(t)$$

Experiment with how you pick the misclassified sample (e.g., train using the same sequence of samples and pick the first one; randomly shuffle the samples after each run through the samples and pick the first misclassified one in the new sequence, etc.). Report the number of updates for convergence. Plot the samples, the target function and the final converged estimated classes on the same figure. Comment on whether the target and the estimated target are close.

c) Repeat everything in b) for datasets of sizes $N = 100$, 1000 and 10000.

d) Summarize your conclusions with respect to the running time as a function of N and the selection method used.

10. Consider the joint density of three random variables a,b,c given by $p(a,b,c) = p(a)p(b \mid a)p(c \mid b)$. It is desired to find the best approximating density $q(a)$ to minimize the Kullback-Leibler (KL) divergence

$$KL(q(a)p(b \mid a) \,||\, p(a,b \mid c)) = \int_a \int_b q(a)p(b \mid a) \ln \frac{q(a)p(b \mid a)}{p(a,b \mid c)} da db$$

Show that the best approximating $q(a)$ is given by

$$q(a) = \frac{p(a)\exp(\int_b p(b \mid a) \ln p(c \mid b) db}{\int_a p(a)\left[\exp(\int_b p(b \mid a) \ln p(c \mid b) db\right] da}$$

Note that when $p(c \mid b) = p(c)$, $q(a)=p(a)$, as it should!

Hint: use the definition of KL divergence and impose the constraint $\int_a q(a)da - 1 = 0$.