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

Problem Set # 1

(Due February 2, 2021)

1. (5 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

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

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

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

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

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

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

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

2. (15 points) (Multivariable Calculus) 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.

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

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

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

$$Note: \nabla_{\underline{w}}(\underline{w}^T\underline{x}_n) = \underline{x}_n; (\underline{w}^T\underline{x}_n)^2 = \underline{w}^T\underline{x}_n\underline{x}_n^T\underline{w} = tr(\underline{w}^T\underline{x}_n\underline{x}_n^T\underline{w}) = tr(\underline{x}_n\underline{x}_n^T\underline{w}\underline{w}^T)$$

$$\nabla^2_{\underline{w}}[tr(\underline{w}^T\underline{x}_n\underline{x}_n^T\underline{w})] = [\frac{\partial^2(\underline{w}^T\underline{x}_n\underline{x}_n^T\underline{w})}{\partial w_i\partial w_j}] = [2x_{ni}x_{nj}] = 2\underline{x}_n\underline{x}_n^T$$

Is the function in (i) 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. (25 points) (ECE 6111 review: Bayes rule): The following are three simple applications of Bayes rule, probability theory and maximum likelihood estimation.

a) (10 points) Consider three random variables A, B, C with the joint distribution P(A,B,C) = P(A) P(B|A) P(C|B) where A has two possible outcomes (0,1), B has three possible outcomes (0,1,2) and z has two possible outcomes (0,1). Let P(A=0)=0.3. The conditional probabilities, P(B|A) and P(C|B) are as shown in the following Tables.

A	B	P(B|A)
0	0	0.30
0	1	0.20
1	0	0.1
1	1	0.4

B	P(C=0|B)
0	0.20
1	0.20
2	0.4

Find P (A, B, C), P (B), P(C) and P (A, C). Is P (A, C) = P (A) P(C)? Is it true in general? What is P (A|C=1) and P (B|C=1).

b) (5 points) An urn contains K balls, of which B are black and (K-B) are white. Fred draws a ball at random from the urn and replaces it, and does this N times.

i. What is the probability distribution of the number of times a black ball is drawn, n_B?

ii. What is E(n_B) and var(n_B)? Give numerical answers for the cases N=5 and N=400, when B=2 and K=10.

iii. Define the fraction $f_B = B/K$. Find E(x) where

$$x = \frac{(n_B - f_B N)^2}{N f_B (1 - f_B)}$$

In the case when $N = 5$ and f_B= 1/5, what is the probability distribution of x? What is the probability that x<1. (Hint: n_B can only take values from 0 to 5).

b) (10 points) Unstable particles are emitted from a source and they decay at a distance x, a real number that has an exponential probability distribution with characteristic length λ (measured in cm). Decay events can be observed only if they occur in a window extending from x = 1 cm to x = 15 cm. N decays are observed at locations $\{x_1, x_2 \,.\,.\,, x_N\}$. Write the likelihood function and derive the necessary condition satisfied by the maximum likelihood estimate of λ?

3. (5 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 $H(\theta)$.

4. (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.
5. (10 points) Using $I(X;Y|Z)=H(X|Z)-H(X|Y,Z)$, compute $I(X;Y|Z)$ when X,Y and Z are scalar Gaussian random variables.
6. (20 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

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

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

   d) Show that

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

   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

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

   g) One nice feature of linear regression is that you can compute prediction of z_i, denoted by $\hat{z}_{-i}$ when the weights are trained on all data except i *without re-computing* the weights. Show that

   $$\hat{\underline{z}}_{-i} = \hat{\underline{w}}_{-i}^{T}\underline{x}_i = \frac{\hat{z}_i - P_{ii}z_i}{1 - P_{ii}};\, P_{ii} = i^{th} \text{ diagonal of } P$$

   h) A method often used to evaluate machine-learning models is the so-called leave-one-out cross-validation (LOCCV). Show that

   $$J_{LOOCV} = \frac{1}{N}\sum_{i=1}^{N}(z_i - \hat{z}_{-i})^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{z_i - \hat{z}_i}{1 - P_{ii}}\right)^2$$

This means that computation of LOOCV cost involves the solution of just one least squares problem. This is not true for nonlinear problems!

i. Show that $E[J_{LOOCV}] = \frac{\sigma^2}{N}\sum_{i=1}^{N}\left(\frac{(1+P_{ii})}{(1-P_{ii})^2}\right)$

7. (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 $z(t) \neq$ sign ($\underline{w}^T(t)\underline{x}(t)$), the update rule is a type of reinforcement learning ("training with a critic") of the form

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

Intuitively, what this means is that if sign ($\underline{w}^T \underline{x}$) =-1 and $z(t)$ =1, we modify $\underline{w}(t)$ so that $\underline{w}^T(t+1)\underline{x}(t)$ increases. On the other hand, if sign ($\underline{w}^T \underline{x}$) =1 and $z(t)$ =-1, we change $\underline{w}(t)$ so that $\underline{w}^T(t+1)\underline{x}(t)$ decreases. 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.