EEL5840 Fundamentals of Machine Learning **Fall 2022**
Midterm Exam **Solutions**

1. (15 points) **Suppose that you are training a linear polynomial regression model of order M for a training set with N data points $\{x_i\}_{i=1}^N$, where $x_i \in \mathbb{R}$, and its corresponding target labels $\{t_i\}_{i=1}^N$. Answer the following questions:**

(a) (5 points) **Write down the mapper function.**

The mapper function for a linear M-th order polynomial regression model can be written as:

$$y(x_i) = \sum_{j=0}^{M} w_j x_i^j = \mathbf{X}\mathbf{x}$$

where $\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ M \end{bmatrix}$ and $\mathbf{X} = \begin{bmatrix} 1 & x_1 & \dots & x_1^M \\ 1 & x_2 & \dots & x_2^M \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & \dots & x_N^M \end{bmatrix}$.

(b) (5 points) **Suppose you want to minimize the absolute error with the Lasso regularizer. Write down the objective function.**

The absolute error objective function with the L1-regularization for the mapper function defined in problem (1) is:

$$J(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{N} |t_i - y(x_i)| + \frac{\lambda}{2} \sum_{j=0}^{M} |w_j|$$

(c) (5 points) **What is the Bayesian interpretation of this objective function? Show and justify your work.**

The Bayesian interpretation of this objective function is equivalent to find the point estimate that maximizes the product of a data likelihood and a prior probability of the parameters $\mathbf{w}$.

$$
\begin{aligned}
&\arg_{\mathbf{w}} \min J(\mathbf{w}) \\
=&\arg_{\mathbf{w}} \max -J(\mathbf{w}) \\
=&\arg_{\mathbf{w}} \max \exp\left(-J(\mathbf{w})\right) \\
=&\arg_{\mathbf{w}} \max \exp\left(-\frac{1}{2}\sum_{i=1}^{N}|t_i - y(x_i)| - \frac{\lambda}{2}\sum_{j=0}^{M}|w_j|\right) \\
=&\arg_{\mathbf{w}} \max \prod_{i=1}^{N}\frac{1}{2}\exp\left(-|t_i - y(x_i)|\right)\prod_{j=0}^{M}\frac{1}{2}\exp\left(-\lambda|w_j|\right) \\
\propto&\arg_{\mathbf{w}} \max \prod_{i=1}^{N}\mathcal{L}\left(t_i|y(x_i),1\right)\prod_{j=0}^{M}\mathcal{L}\left(w_j|0,1/\lambda\right) \\
=&\arg_{\mathbf{w}} \max \mathcal{L}\left(\mathbf{t}|\mathbf{y},\mathbf{1}\right)\mathcal{L}\left(\mathbf{w}|0,1/\lambda\right)
\end{aligned}
$$

As we see above, the data likelihood modeling the target variable $\mathbf{t}$ is described by a Laplace distribution with mean $\mathbf{y}(\mathbf{x})$ and scale 1. And, the prior probability modeling the parameters $\mathbf{w}$ also follow a Laplace distribution with mean 0 and scale $1/\lambda$.

2. (10 points) **Consider a linear basis function regression model for a multivariate target variable t having a multivariate Gaussian distribution of the form**

$$p(\mathbf{t}|\mathbf{W},\mathbf{\Sigma}) = \mathcal{N}\left(\mathbf{t}|y(x,\mathbf{W}),\mathbf{\Sigma}\right)$$

where

$$y(\mathbf{x},\mathbf{W}) = \mathbf{XW} \in \mathbb{R}^{N\times M}$$

together with a training data set comprising input basis vectors $\phi(x_n) \in \mathbb{R}^{D\times 1}$ and corresponding target vectors t_n, $t_n \in \mathbb{R}^M$, with $n = 1,\ldots,N$. N corresponds to the number of samples, M corresponds to the number of target values, and D the dimensionality of the feature space. Moreover, $\mathbf{W} \in \mathbb{R}^{D\times M}$ and $\mathbf{X} = \begin{bmatrix}\phi(x_1)^T \\ \phi(x_2)^T \\ \vdots \\ \phi(x_N)^T\end{bmatrix} \in \mathbb{R}^{N\times D}$.

Solve for the maximum likelihood (MLE) solution for the parameter matrix W, under the assumption that the Gaussian distribution has an isotropic covariance matrix. Show all your work.

As instructed, let's consider $\Sigma = \sigma^2\mathbf{I}$. Under this assumption, the determinant $|\sigma| = (\sigma^2)^D$, and its inverse is $\Sigma^{-1} = (\sigma^2)^{-1}\mathbf{I}$.

The MLE solution for the parameter $\mathbf{W}$ can be found by finding the point estimate that maximizes the log-data likelihood $\mathcal{L}$ which is defined as:

$$\begin{aligned}
\mathcal{L} &= \ln\left(\mathcal{L}^0\right) \\
&= \ln\left[\prod_{i=1}^{N} \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(t_i - y_i)^T\Sigma^{-1}(t_i - y_i)\right)\right] \\
&= \ln\left[\prod_{i=1}^{N} \frac{1}{(2\pi)^{D/2}(\sigma^2)^{D/2}} \exp\left(-\frac{1}{2\sigma^2}(t_i - y_i)^T(t_i - y_i)\right)\right] \\
&= \sum_{i=1}^{N}\left[-\frac{D}{2}\ln(2\pi) - \frac{D}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(t_i - y_i)^T(t_i - y_i)\right] \\
&= \sum_{i=1}^{N}\left[-\frac{D}{2}\ln(2\pi) - \frac{D}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(t_i - \phi^T(x_i)\mathbf{W})^T(t_i - \phi^T(x_i)\mathbf{W})\right] \\
&= \sum_{i=1}^{N}\left[-\frac{D}{2}\ln(2\pi) - \frac{D}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(t_i^Tt_i - t_i^T\phi^T(x_i)\mathbf{W} - \mathbf{w}\phi(x_i)t_i + \mathbf{W}\phi(x_i)\phi^T(x_i)\mathbf{W}^T)\right]
\end{aligned}$$

Now, we find the solution for $\mathbf{w}$:

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\mathbf{W}} = 0 &\iff \sum_{i=1}^{N} -\frac{1}{2\sigma^2}\left[-2t_i^T\phi^T(x_i) + 2\mathbf{w}^T\phi(x_i)\phi^T(x_i)\right] = 0 \\
&\iff \sum_{i=1}^{N}\phi(x_i)t_i = \sum_{i=1}^{N}\phi(x_i)\phi^T(x_i)\mathbf{w} \\
&\iff \mathbf{X}^T\mathbf{t} = \mathbf{X}^T\mathbf{X}\mathbf{W} \\
&\iff \mathbf{W}_{\text{MLE}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}
\end{aligned}$$

As expected, the MLe solution for $\mathbf{W}$ is $\mathbf{W}_{\text{MLE}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$.

3. (10 points) **Consider the data likelihood from problem 2 with a multivariate Gaussian prior distribution on the parameters W with mean zero and isotropic covariance $\Sigma_0 = \beta\mathbf{I}$, that is,**

$$p(\mathbf{W}) = \mathcal{N}\left(\mathbf{W}|0, \beta\mathbf{I}\right)$$

where β is a user-defined constant. Solve for the maximum a posteriori (MAP) solution for the parameter matrix $\mathbf{W}$. Show all your work.

The MAP solution for the parameter $\mathbf{W}$ can be found by finding the point estimate that maximizes the log-data likelihood $\mathcal{L}$ which is defined as:

$$\begin{aligned}
\mathcal{L} =& \ln \mathcal{N}\left(\mathbf{t}|y(x,\mathbf{W}),\sigma^2\mathbf{I}\right)\mathcal{N}\left(\mathbf{W}|0,\beta\mathbf{I}\right)\\
=& \ln\left[\prod_{i=1}^{N}\frac{1}{(2\pi)^{D/2}(\sigma^2)^{D/2}}\exp\left(-\frac{1}{2\sigma^2}(t_i-y_i)^T(t_i-y_i)\right)\right]\frac{1}{(2\pi)^{D/2}|\beta\mathbf{I}|^{1/2}}\exp\left(-\frac{1}{2\beta}\mathbf{W}^T\mathbf{W}\right)\\
=& \sum_{i=1}^{N}\left[-\frac{D}{2}\ln(2\pi)-\frac{D}{2}\ln(\sigma^2)-\frac{1}{2\sigma^2}(t_i^Tt_i-t_i^T\phi^T(x_i)\mathbf{W}-\mathbf{w}\phi(x_i)t_i+\mathbf{W}\phi(x_i)\phi^T(x_i)\mathbf{W}^T)\right]\\
&-\frac{D}{2}\ln(2\pi)-\frac{D}{2}\ln(\beta)-\frac{1}{2\beta}\mathbf{W}^T\mathbf{W}
\end{aligned}$$

Now, we find the solution for $\mathbf{w}$:

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\mathbf{W}}=0 \iff& \sum_{i=1}^{N}-\frac{1}{2\sigma^2}\left[-2t_i^T\phi^T(x_i)+2\mathbf{w}^T\phi(x_i)\phi^T(x_i)\right]-\frac{1}{\beta}\mathbf{W}^T=0\\
\iff& \frac{1}{\sigma^2}\left(\mathbf{t}^T\mathbf{X}-\mathbf{W}^T\mathbf{X}^T\mathbf{X}\right)-\frac{1}{\beta}\mathbf{W}^T=0\\
\iff& \beta\mathbf{t}^T\mathbf{X}=\beta\mathbf{W}^T\mathbf{X}^T\mathbf{X}+\sigma^2\mathbf{W}^T\\
\iff& \beta\mathbf{X}^T\mathbf{t}=\beta\mathbf{X}\mathbf{X}^T\mathbf{W}+\sigma^2\mathbf{W}\\
\iff& \mathbf{X}^T\mathbf{t}=\left(\mathbf{X}\mathbf{X}^T+\frac{\sigma^2}{\beta}\right)\mathbf{W}\\
\iff& \mathbf{W}_{\text{MAP}}=\left(\mathbf{X}\mathbf{X}^T+\frac{\sigma^2}{\beta}\right)^{-1}\mathbf{X}^T\mathbf{t}
\end{aligned}$$

4. (10 points) **Consider the mapper function and the optimization problem presented in problem 3. Answer the following questions:**

   (a) (3 points) **What are the hyperparameters? Please specify.**

   The hyperparameters are the polynomial model order M and the regularization term β.

   (b) (3 points) **In practice, how do you optimize for the hyperparameters?**

   In practice, we use a cross-validation scheme to identify which value best works for the hyperparameter set $\Theta=\{M,\lambda\}$ such that the perfomance measure is optimized for both training and validation sets.

   There are many cross-validation strategies we can use, including: k-fold cross-validation with or without stratified partition, Bootstrap sampling or $k\times 2$ cross-validation.

(c) (4 points) **Suppose that you have a small training set. In practice, will this information change the strategy you use to optimize for the hyper-parameters? If yes, elaborate on the strategy that you will use. If not, justify why not.**

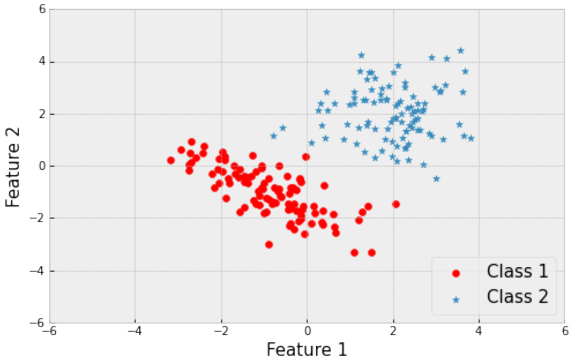
(A wide range of answers are accepted for this question.)

Utilizing k-fold cross-validation will impose a challenge as the dataset is small. If the number of folds is large, the validation sets will be small and likely not going to be able to access the full generalization ability of the model.

For a small dataset with heterogenous samples, it is requested that we use Bootstrap resampling or k-fold cross-validation with small value of k. An alternative is to use the leave-one-out scheme of cross-validation.

It is also always a good idea to regularize the parameters of the model in order to avoid overfitting which is likely to happen for small datasets.

5. (10 points) **Consider the dataset depicted below composed of two classes (class 1 and class 2) in a 2-dimensional feature space. Each class has 100 samples. This is what the data looks like:**



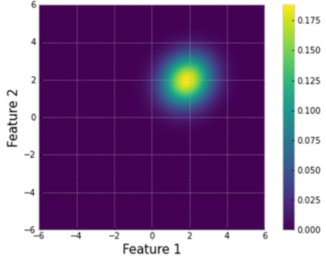
Answer the following questions:

(a) (3 points) **The four plots below (A, B, C and D) represent contours plots of the data likelihood as modeled with a Gaussian distribution. From the four plots, select two plots: one that represents the data likelihood for class 1, and another that represents the data likelihood for class 2. Justify your answer.**

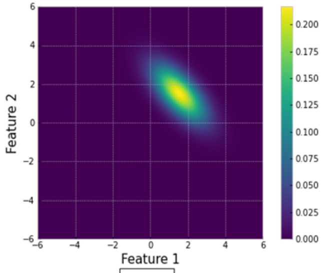
- Class 1: the data has approximately a sample mean of $[-1, -1]$ and an elliptical covariance matrix with negative correlation between features 1 and 2. The best contour fitting this description is surface D.
- Class 2: the data has approximately a sample mean of $[2, 2]$ and a circular covariance matrix with approximately zero correlation between features 1 and 2. The best contour fitting this description is surface A.

(b) (4 points) **In order to train the Naïve Bayes Classifier for this dataset, what other information do you need obtain in order to be able to make predictions? Provide your estimates below.**

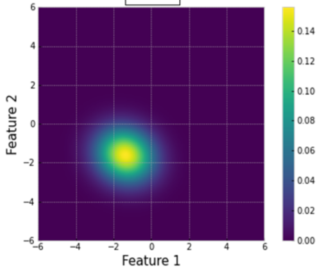
A



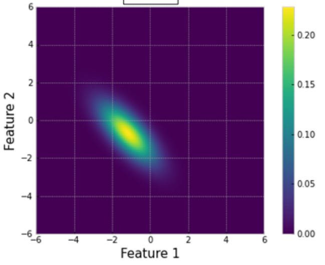
B



C



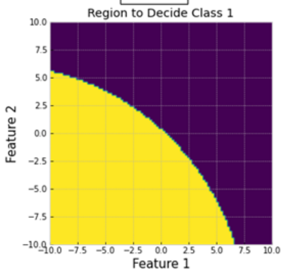
D



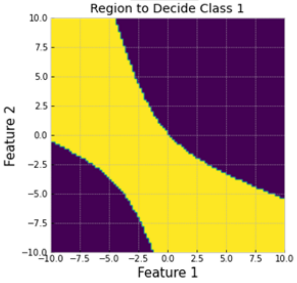
In addition to the data likelihoods for class 1 and class 2, we also need to define the prior probability for each class. Namely, we can consider $P(C_k) = \frac{N_k}{N}$ where N_k is the number of samples assigned to class C_k, N is the total number of samples and $k = \{1, 2\}$.

(c) (3 points) **The three plots below (i, ii and iii) represent the decision surface for deciding class 1 (red class). Based on data likelihoods you selected in part 5.1, which of the following plots corresponds to the decision surface for deciding class 1? Justify your answer.**

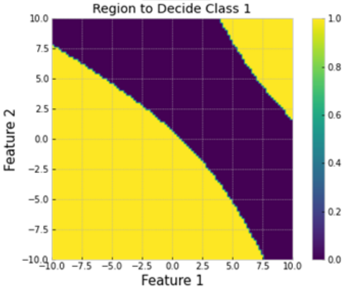
i



ii



iii



Decision surface ii corresponds to the decision surface for deciding class 1, because (1) the sample mean is included in the region to decide class 1, and (2) the decision surface shows that the data likelihood covers quadrants 2 and 4 to accommodate the large uncertainty along that direction.

6. (20 points) **Suppose you have a training set with N data points $\{x_i\}_{i=1}^N$, where $x_i \in \mathbb{R}^+$ (set of positive real numbers). Assume the samples are independent**

and identically distributed (i.i.d.), and each sample is drawn from a Gamma random variable with probability density function:

$$p(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

where $\alpha, \beta > 0$**.**

Moreover, consider another Gamma density as the prior probability on the hyperparameter β**,**

$$p(\beta|a,b) = \frac{b^a}{\Gamma(a)} \beta^{a-1} e^{-b\beta}$$

where $a, b > 0$**.**

Answer the following questions:

(a) (5 points) **Derive the maximum likelihood estimate (MLE) for the parameter** β**. Show your work.**

The observed data likelihood is:

$$\mathcal{L}^0 = \prod_{i=1}^{N} \frac{\beta^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\beta x_i}$$

The log-likelihood is given by:

$$\begin{aligned}\mathcal{L} &= \ln \mathcal{L}^0 \\ &= \sum_{i=1}^{N} \left(\alpha \ln(\beta) - \ln(\Gamma(\alpha)) + (\alpha-1)\ln(x_i) - \beta x_i\right)\end{aligned}$$

We can now find the MLE estimation for β:

$$\frac{\partial \mathcal{L}}{\partial \lambda} = 0 \iff \sum_{i=1}^{N} \left(\alpha \frac{1}{\beta} - x_i\right) = 0 \iff \frac{\alpha N}{\beta} = \sum_{i=1}^{N} x_i \iff \beta = \frac{\alpha N}{\sum_{i=1}^{N} x_i}$$

(b) (5 points) **Derive the maximum a posteriori (MAP) estimate for the parameter** β**. show your work.**

The observed data likelihood for MAP is:

$$\begin{aligned}
\mathcal{L}^0 &= \left(\prod_{i=1}^{N} \frac{\beta^{\alpha}}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\beta x_i}\right) \frac{b^a}{\Gamma(a)} \beta^{a-1} e^{-b\beta} \\
&\propto \left(\prod_{i=1}^{N} \beta^{\alpha} e^{-\beta x_i}\right) \beta^{a-1} e^{-b\beta} \\
&= \beta^{\sum_{i=1}^{N} \alpha} e^{-\beta \sum_{i=1}^{N} x_i} \beta^{a-1} e^{-b\beta} \\
&= \beta^{N\alpha+a-1} e^{-\beta\left(\sum_{i=1}^{N} x_i + b\right)}
\end{aligned}$$

We can now find the MAP estimation for β:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \lambda} = 0 &\iff (N\alpha + a - 1)\ln(\beta) - \beta\left(\sum_{i=1}^{N} x_i + b\right) = 0 \\
&\iff \beta = \frac{N\alpha + a - 1}{\sum_{i=1}^{N} x_i + b}
\end{aligned}$$

(c) (5 points) **Is the Gamma distribution a conjugate prior for the parameter β, of the Gamma data likelihood distribution? Why or why not?**

Yes, the Gamma distribution forms a conjugate prior because the shape of the posterior probability is proportionally equal to the prior probability.

(d) (5 points) **Suppose you would like to update the Gamma prior distribution and the MAP point estimation in an online fashion, as you obtain more data. Write the pseudo-code for the online update of the prior parameters. In your answer, specify the new values for the parameters of the prior.**

The pseudo-code for online update of the prior is as follows:

1. Start at iteration $t = 0$. Initialize the prior parameters $a^{(t)}$ and $b^{(t)}$.
2. Compute the parameter estimation for the current prior probability:

$$\beta_{\text{MAP}} = \frac{N\alpha + a^{(t)} - 1}{\sum_{i=1}^{N} x_i + b^{(t)}}$$

3. Update the prior parameters

$$\begin{aligned}
a^{(t+1)} &\leftarrow a^{(t)} + N\alpha \\
b^{(t+1)} &\leftarrow b^{(t)} + \sum_{i=1}^{N} x_i
\end{aligned}$$

4. Increment iteration counter

$$t \leftarrow t + 1$$

7. (25 points) **Consider a training set containing positive real numbers ($x \in \mathbb{R}^+$) for 2 classes, C_0 and C_1. The training set has 400 samples for class C_0 and 100 for C_1.**

   Suppose that you have reason to believe that samples belonging from C_0 are drawn from an Exponential random variable with parameter $\lambda > 0$, and samples belonging to C_1 are drawn from a Gamma random variable with parameters $\alpha > 0$ and $\beta > 0$. In other words:

$$p(x|C_0) = \lambda e^{-\lambda x}$$
$$p(x|C_1) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

   where $\Gamma(x) = (x-1)!$. Answer the following questions:

   (a) (5 points) **For a given training set $\{(x_i, t_i)\}_{i=1}^N$ with $x_i \in \mathbb{R}^+$ and $t_i \in \{0, 1\}$, how would you determine the parameters for each classes' data likelihood (λ, α and β)? Explain in words a list of steps for finding the point estimates. (No derivations.)**

   The parameters λ, α and β can be estimated as point estimates of the data likelihood either using MLE, a completely data-driven approach without any regularization of the parameter values; or using MAP, the Maximum A Posteriori approach which will induce a constraint on the parameter values.

   (b) (5 points) **Can you regularize your point estimates for λ, α and β? If yes, clearly explain how you would proceed. If not, explain why not.**

   Yes, as mentioned in the previous question, λ, α and β can be regularized. In order to do this, we would have to use the MAP approach.

   (c) (10 points) **For next two parts, consider $\lambda = 2$, $\alpha = 2$ and $\beta = 1$. Consider the test point $x = 3$. Which class (C_0 or C_1) will it be assigned to? Show your work.**

   The prior probability for each class can be estimated as its relative frequency in the training set, namely

$$p(C_1) = \frac{400}{500} = \frac{4}{5} \text{ and } p(C_2) = \frac{100}{500} = \frac{1}{5}$$

   For the provided parameter values, the data likelihood for class 1 and class 2 are defined as:

$$p(x|C_0) = 2e^{-2x}$$
$$p(x|C_1) = \frac{1^2}{\Gamma(2)} x^{2-1} e^{-1x} = xe^{-x}$$

For the test point $x = 3$, we compute the posterior probability with

$$P(C_k|x=3) = \frac{p(x=3|C_k)p(C_k)}{p(x=3)} = \frac{p(x=3|C_k)p(C_k)}{\sum_{j=1}^{2} p(x=3|C_j)p(C_j)}$$

We now find:

$$\begin{aligned} P(C_0|x=3) &= \frac{2e^{-6}\frac{4}{5}}{2e^{-6}\frac{4}{5} + 3e^{-3}\frac{1}{5}} \approx 0.117 \\ P(C_1|x=3) &= \frac{3e^{-3}\frac{1}{5}}{2e^{-6}\frac{4}{5} + 3e^{-3}\frac{1}{5}} \approx 0.883 \end{aligned}$$

Since $P(C_1|x=3) > P(C_0|x=3)$ then the test point $x = 3$ is assigned to class C_1.

(d) (5 points) **For a given test sample x, provide an equation that will determine all cases in each x will be assigned to C_0. Show your work.**

A point x is assigned to class C_0 if $P(C_0|x) > P(C_1|x)$:

$$\begin{aligned} P(C_0|x) &> P(C_1|x) \\ \frac{p(x|C_0)p(C_0)}{p(x)} &> \frac{p(x|C_1)p(C_1)}{p(x)} \\ \frac{p(x|C_0)}{p(x|C_1)} &> \frac{p(C_1)}{p(C_0)} \\ \frac{2e^{-2x}}{xe^{-x}} &> \frac{1/5}{4/5} \\ \frac{e^{-x}}{x} &> \frac{1}{8} \end{aligned}$$

$\therefore x \in C_0$ if $\frac{e^{-x}}{x} > \frac{1}{8}$.