EEL6825 Pattern Recognition and Intelligent Systems **Spring 2023**
Midterm Exam **Solutions**

1. (10 points) **In class, we studied the sample complexity bounds for a hypothesis class (or mapper function), $\mathcal{H}$, in a supervised learning system. In both realizable and agnostic cases, the sample complexity bound is a function of the cardinality of the hypothesis class, $|\mathcal{H}|$. The larger $|\mathcal{H}|$, the more training examples are needed to learn a concept.**

   In practice, how would you pick a hypothesis class so that its cardinality is as small as possible? Provide 2 strategies and explain why.

   One strategy would be to select the set of $\mathcal{H}$ with simple representations (for example, linear classifiers or regression mapper functions) - the Occam's Razor strategy. However, this constraint on simplicity of class $\mathcal{H}$ may produce a large empirical error.

   Other strategies include: regularization and cross-validation. In regularization, we apply a constraint on the values of the parameters of the mapper function (e.g. with ridge or lasso penalties), thus allowing for more complex classes to be selected and, at the same time, reducing the cardinality of $\mathcal{H}$. In cross-validation, we utilize a validation set to learn the set of hypothesis that fit the data but also generalize to unseen examples. The use of a cross-validation strategy will impose a limit on the complexity of possible functions of $\mathcal{H}$ thus reducing its cardinality.

2. (10 points) **Under the non-parametric density estimation topic, we talked about Kernel Density Estimation and K-Nearest Neighbor Density Estimation. Answer the following questions:**

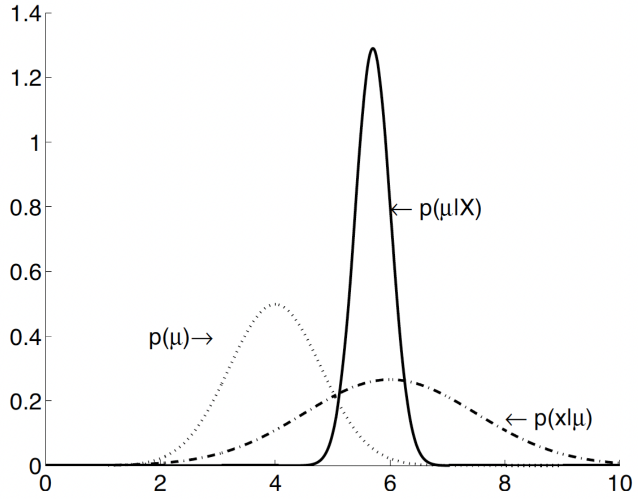
   (a) (5 points) **What is the main difference between these two approaches?**
   In the KDE estimation of the density function, the volume around the points x was considered fixed (h_n) and the number of points k_n, falling inside the volume, was left to vary randomly from point to point. In the KNN estimation of the density function, the number of points $k_n = k$ will be fixed, and the size of the volume around x will be adjusted each time, to include k points. Thus, in low-density areas the volume will be large and in high-density areas it will be small.

   (b) (5 points) **What are the hyperparameters for each approach?** In KDE, the hyperparameters include: choice of kernel function (Gaussian, Exponential, etc.) and the kernel bandwidth.

   In KNN, the hyperparamters are the number of points k.

3. (15 points) **Consider now a parametric density estimation approach with Maximum A Posteriori (MAP). Suppose you have a training set $\{x_i\}_{i=1}^N$, where $x_i \in \mathbb{R}$. Assume the samples are independent and identically distributed (i.i.d.), and each sample is drawn from a (univariate) Gaussian density function with known variance σ^2, i.e. $P(x|\mu) \sim \mathcal{N}(\mu, \sigma^2)$.**

   Consider a prior belief on the parameter μ to be modeled according to another (univariate) Gaussian, $P(\mu) \sim \mathcal{N}(\mu_0, \sigma_0^2)$.



If we let $P(x|\mu) \sim \mathcal{N}(\mu = 6, \sigma^2 = 1.5^2)$ and $P(\mu) \sim \mathcal{N}(\mu_0 = 4, \sigma_0^2 = 0.8^2)$, then the resulting posterior probability on μ is $P(\mu|x) \sim \mathcal{N}(5.7, 0.3^2)$. The figure above illustrates all distributions involved.

Answer the following questions:

(a) (5 points) **Based on this plot, what is the numerical value of μ that will be selected by MAP?**

The value for μ that will be selected is the one that maximizes the posterior probability. Since the posterior is a Gaussian distribution, the point with the largest density value is the mean which is given at $\mu = 5.7$. We can also use the plot to visually confirm this value.

(b) (5 points) **How would the solution for μ change if σ_0 is very small?**

If the variance of the prior distribution, σ_0, is **very small**, then the prior encodes a **strong prior belief**. Then the value of μ will be strongly influenced by this strong prior and will be approximately the mean of the prior probability, $\mu = 4$.

(c) (5 points) **How would the solution for μ change if σ_0 is very large?**

If the variance of the prior distribution, σ_0, is **very large**, then the prior encodes a **weak prior belief** (converging to a "flat" distribution). Then the value of μ will be strongly influenced by the data and will be approximately the mean of the data likelihood, $\mu = 6$.

4. (25 points) **Suppose you have a training set with N data points $\{x_i\}_{i=1}^N$ with $x_i \in \mathbb{R}$. Assume the samples are independent and identically distributed (i.i.d.), and each sample is drawn from a (univariate) Gaussian density function with known mean μ:**

$$P(x|\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad \text{(with known } \mu)$$

To simplify notation, take the variance $\sigma^2 = \nu$:

$$P(x|\nu) = \frac{1}{\sqrt{2\pi\nu}} \exp\left(-\frac{(x-\mu)^2}{2\nu}\right) \quad \text{(with known } \mu\text{)}$$

Moreover, consider the Inverse-gamma density function as the prior probability on the variance parameter, ν,

$$P(\nu) = \frac{\beta^\alpha}{\Gamma(\alpha)} \nu^{-\alpha-1} \exp\left(-\frac{\beta}{\nu}\right)$$

where $\alpha > 0$ and $\beta > 0$. Answer the following questions:

(a) (10 points) **Derive the maximum likelihood estimate (MLE) for the variance parameter ν. Show your work.**

The observed data likelihood for MLE is:

$$\mathcal{L}^0 = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\nu}} \exp\left(-\frac{(x_i-\mu)^2}{2\nu}\right)$$

The log-likelihood is given by:

$$\begin{aligned}\mathcal{L} = \ln \mathcal{L}^0 &= \sum_{i=1}^{N} \left(-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\nu) - \frac{(x_i-\mu)^2}{2\nu}\right) \\ &= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\nu) - \frac{1}{2\nu}\sum_{i=1}^{N}(x_i-\mu)^2\end{aligned}$$

We can now find the MLE estimation for ν:

$$\frac{\partial \mathcal{L}}{\partial \nu} = 0 \iff -\frac{N}{2\nu} + \frac{2\sum_{i=1}^{N}(x_i-\mu)^2}{(2\nu)^2} \iff \nu_{\text{MLE}} = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$$

(b) (10 points) **Derive the maximum a posteriori (MAP) estimate for the variance parameter ν. Show your work.**

The observed data likelihood for MAP is:

$$\begin{aligned}\mathcal{L}^0 &= \left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\nu}} \exp\left(-\frac{(x_i-\mu)^2}{2\nu}\right)\right) \frac{\beta^\alpha}{\Gamma(\alpha)} \nu^{-\alpha-1} \exp\left(-\frac{\beta}{\nu}\right) \\ &= (2\pi)^{-N/2} \frac{\beta^\alpha}{\Gamma(\alpha)} \nu^{-\frac{N}{2}-\alpha-1} \exp\left(-\frac{1}{\nu}\left(\frac{(x_i-\mu)^2}{2} + \beta\right)\right) \\ &\propto \nu^{-\left(\frac{N}{2}+\alpha\right)-1} \exp\left(-\frac{1}{\nu}\left(\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2} + \beta\right)\right)\end{aligned}$$

The log-likelihood is given by:

$$\mathcal{L} = \left(-\frac{N}{2} - \alpha - 1\right)\ln(\nu) - \frac{1}{\nu}\left(\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2} + \beta\right)$$

We can now find the MAP estimation for ν:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \nu} = 0 &\iff \left(-\frac{N}{2} - \alpha - 1\right)\frac{1}{\nu} + \frac{2\left(\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2} + \beta\right)}{\nu^2} \\ &\iff \frac{2\left(\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2} + \beta\right)}{\nu} = \frac{N}{2} + \alpha + 1 \\ &\iff \nu_{\text{MAP}} = \frac{\sum_{i=1}^{N}(x_i-\mu)^2 + 2\beta}{N/2 + \alpha + 1}\end{aligned}$$

(c) (5 points) **Does the Gaussian distribution and the Inverse-gamma prior on the parameter ν have a conjugate prior relationship? Why or why not?**

If yes, write down the pseudo-code for the online update of the prior parameters (α and β) and specify how the prior parameters will be updated.

Yes, the Gaussian distribution and the Inverse-gamma prior on the parameter ν have a conjugate prior relationship because the shape of the posterior probability is proportionally equal to the prior probability.

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

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

$$\nu_{\text{MAP}} = \frac{\sum_{i=1}^{N}(x_i-\mu)^2 + 2\beta}{N/2 + \alpha + 1}$$

3. Update the prior parameters

$$\begin{aligned}\alpha^{(t+1)} &\leftarrow \alpha^{(t)} + \frac{N}{2} \\ \beta^{(t+1)} &\leftarrow \beta^{(t)} + \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2}\end{aligned}$$

4. Increment iteration counter

$$t \leftarrow t + 1$$

5. (20 points) **Consider a binary classification task (with classes C_0 and C_1). The prior probability for C_i is $P(C_i) = p_i$. Each class is modeled with a bivariate Gaussian distribution,**

$$P(\mathbf{x}|C_1) \sim \mathcal{N}\left(\mu_1 = \begin{bmatrix}1\\0\end{bmatrix}, \Sigma_1 = \mathbf{I}\right) \quad \text{and} \quad P(\mathbf{x}|C_0) \sim \mathcal{N}\left(\mu_2 = \begin{bmatrix}0\\-1\end{bmatrix}, \Sigma_2 = \mathbf{I}\right)$$

where $\mathbf{x} = \begin{bmatrix}x_1\\x_2\end{bmatrix}$ and $\mathbf{I}$ is the 2×2 identity matrix.

(a) (10 points) **Let $g_i(\mathbf{x})$ be the discriminant function for class C_i. Then, let $g(\mathbf{x}) = g_1(\mathbf{x}) - g_0(\mathbf{x})$ be the Naïve Bayes discriminant function. Find $g(\mathbf{x})$. Show all your work.**

To compute the data likelihoods, we need to determine the inverse of the covariance matrix. Since they are both the identity matrix $\Sigma_1 = \Sigma_2 = \mathbf{I}$, we know that $\Sigma_1^{-1} = \Sigma_2^{-1} = \mathbf{I}^{-1} = \mathbf{I}$.

The discriminant function $g_i(x)$, $i = 1, 2$ is defined as $P(x|C_i)P(C_i)$. We can simplify it by taking $g_i(x) = \ln(P(x|C_i)P(C_i))$. Note that $x = [x_1, x_2]^T$.We have:

$$\begin{aligned}
g_1(x) &= \ln\left(p_1 \times \frac{1}{(2\pi)^{2/2}|\Sigma_1|^{1/2}} \exp\left\{-\frac{1}{2}\begin{bmatrix}x_1 - 1\\x_2\end{bmatrix}^T \Sigma_1^{-1}\begin{bmatrix}x_1 - 1\\x_2\end{bmatrix}\right\}\right)\\
&= \ln(p_1) - \ln(2\pi) - \frac{1}{2}((x_1 - 1)^2 + x_2^2)\\
&= \ln(p_1) - \ln(2\pi) - \frac{1}{2}(x_1^2 - 2x_1 + 1 + x_2^2)
\end{aligned}$$

and

$$\begin{aligned}
g_0(x) &= \ln\left(p_0 \times \frac{1}{(2\pi)^{2/2}|\Sigma_2|^{1/2}} \exp\left\{-\frac{1}{2}\begin{bmatrix}x_1\\x_2 + 1\end{bmatrix}^T \Sigma_2^{-1}\begin{bmatrix}x_1\\x_2 + 1\end{bmatrix}\right\}\right)\\
&= \ln(p_0) - \ln(2\pi) - \frac{1}{2}(x_1^2 + (x_2 + 1)^2)\\
&= \ln(p_0) - \ln(2\pi) - \frac{1}{2}(x_1^2 + x_2^2 + 2x_2 + 1)
\end{aligned}$$

Now, let $g(x) = g_1(x) - g_0(x)$, we find:

$$\begin{aligned}
g(x) &= -\frac{1}{2}(x_1^2 - 2x_1 + 1 + x_2^2) + \frac{1}{2}(x_1^2 + x_2^2 + 2x_2 + 1) + \ln\left(\frac{p_1}{p_0}\right)\\
&= x_1 + x_2 + \ln\left(\frac{p_1}{p_0}\right)
\end{aligned}$$

Thus, we decide C_1 (label 1) if $g(x) > 0 \iff x_1 + x_2 + \ln\left(\frac{p_1}{p_0}\right) > 0$.

(b) (10 points) **In this part, consider $P(C_1) = p_1 = \frac{1}{5}$.**
For each sample (x_1, x_2), the table below includes the true label assignment (where C_1 is marked as 1, and C_0 is marked as 0). Use the Naïve Bayes classifier from part (a) to predict the labels for each sample. Show your work. (Note: $\ln(1/4) \approx -1.3862$.**)**

Since $p_1 = \frac{1}{5}$, then $p_0 = 1 - p_1 = \frac{4}{5}$. Thus $\ln\left(\frac{p_1}{p_0}\right) = \ln\left(\frac{1}{4}\right) \approx -1.3862$.

(x_1, x_2)	t	Classifier Prediction
$(0, 0)$	0	$0 + 0 - 1.3862 < 0$, then 0
$(1, 0)$	1	$1 + 0 - 1.3862 < 0$, then 0
$(-1, 0)$	1	$-1 + 0 - 1.3862 < 0$, then 0
$(0.5, 0.5)$	1	$0.5 + 0.5 - 1.3862 < 0$, then 0
$(2, -0.5)$	0	$2 - 0.5 - 1.3862 > 0$, then 1
$(2, 2)$	1	$2 + 2 - 1.3862 > 0$, then 1
$(-1, 2)$	1	$-1 + 2 - 1.3862 < 0$, then 0
$(0.25, 0)$	0	$0.25 + 0 - 1.3862 < 0$, then 0
$(0, 0.25)$	0	$0 + 0.25 - 1.3862 < 0$, then 0
$(-1, -1)$	0	$-1 - 1 - 1.3862 < 0$, then 0

6. (10 points) **Define action α_i as the decision to assign the input sample x to class C_i and λ_{ik} the *loss* incurred for taking action α_i when the input actually belongs to C_k.**

In a two-class, three-action problem, consider the following loss function,

	α_1	α_2	α_r
C_1	$\lambda_{11} = 0$	$\lambda_{21} = 5$	$\lambda_{31} = \lambda$
C_2	$\lambda_{12} = 10$	$\lambda_{22} = 0$	$\lambda_{32} = \lambda$

where α_r is the reject action. In what conditions will the system choose α_1, α_2 and α_r? Show your work.

An action will be decided if its risk is the smallest. Based off of the loss function table, the risk for each action is as follows:

$$\begin{aligned} R(\alpha_1|x) &= 10P(C_2|x) \\ R(\alpha_2|x) &= 5P(C_1|x) \\ R(\alpha_r|x) &= \lambda \end{aligned}$$

The optimal decision rule is:

$$\begin{gathered} \text{Choose } \; C_i \; \text{ if } \; R(\alpha_i|\mathbf{x}) < R(\alpha_k|\mathbf{x}) \; \text{ for all } k \neq i \text{ and } R(\alpha_i|\mathbf{x}) < \lambda \\ \text{Reject} \quad \text{otherwise} \end{gathered}$$

Thus,

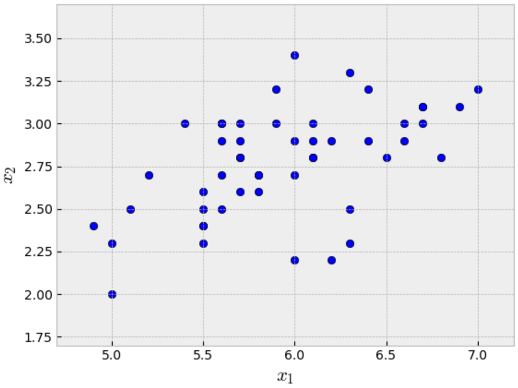
$$\begin{aligned} \text{Choose} \quad \alpha_1 : \;& 10P(C_2|x) < \lambda \iff 10(1 - P(C_1|x)) < \lambda \iff P(C_1|x) > 1 - \frac{\lambda}{10} \\ \text{Choose} \quad \alpha_2 : \;& 5P(C_1|x) < \lambda \iff P(C_1|x) < \frac{\lambda}{5} \\ \text{Reject} \quad : \;& \frac{\lambda}{5} < P(C_1|x) < 1 - \frac{\lambda}{10} \end{aligned}$$

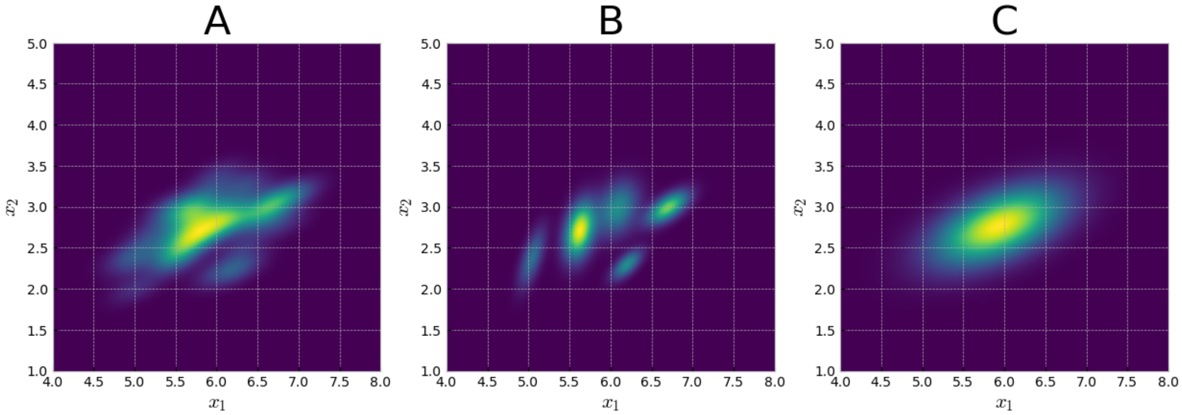
7. (10 points) **Consider the following 2-dimensional dataset:**

Suppose you run three (3) different approaches for density estimation and obtain the following results:

From the different approaches for density estimation seen in lecture, identify which approach was utilized to produce results A, B and C. Justify your answer.

The most noticeable characteristics between these 3 density estimations is that result C appears have a single mode and, results A and B, appear to have multiple modes.





The result in C depicts the contour of a Gaussian distribution with a full covariance matrix and positive correlation between features x_1 and x_2. It is clear that it is using a parametric distribution, namely Gaussian distribution, to describe the dataset. The possible solutions that could have produced this density estimation are: MLE or MAP. (Note that we could have selected GMM with 1 component - this would be a redundant approach since GMM with 1 component is equivalent to using MLE, which is computationally cheaper).

For results A and B, as stated earlier, both have multiple modes. This is indicative that GMMs, KDE or KNN density estimation was used. In result B, it appears that there exist 5 Gaussian-shaped contours with minimal to no blurriness. This is indicative that a GMM with 5 components was used for density estimation.

In result A, the ”blurriness” between densely populated areas (areas with high density) are evidently produced by KDE with a relatively large kernel bandwidth. The large kernel bandwidth is averaging the density around each data sample.