Please print in pen:

Waterloo Student ID Number:

WatIAM/Quest Login Userid:

UNIVERSITY OF WATERLOO

Examination
Midterm
Fall 2024
CS 480/680

Times: Tuesday 2024-10-29 at 08:30 to 09:50

Duration: 1 hour 20 minutes (80 minutes)

Exam ID: 5971966

Sections: CS 480 LEC 001,002
CS 680 LEC 001,002

Instructors: Kathryn Simone

Special Materials

Candidates may bring only the listed aids.

- Calculator - Non-Programmable
- Study Notes - Double-Sided 8.5x11 (One double-sided page allowed)

Instructions:

- This exam contains 15 pages (including this cover page) and 5 questions. Total of points is 100.
- All answers should be placed in the spaces given. If you need more space to complete an answer, you may use the blank pages at the end.
- Cheating is an academic offense. Your signature on this exam indicates that you understand and agree to the University's policies regarding cheating on exams.

Name: Solutions

Signature: ______________________

Distribution of Marks

Question	Points	Score
1	40	
2	10	
3	10	
4	10	
5	30	
Total:	100	

+5 correct choice
0 otherwise

Part I: Multiple Choice

1. (a) (5 points) Suppose you are collaborating with a colleague to train a kernel SVM on a training dataset of $n = 2$ samples and $m = 1$ features. They provided to you the following $n \times n$ Kernel matrix so that you could then solve the dual of the SVM problem:

$$\boldsymbol{K} = \begin{bmatrix} 1 & 2 \\ 2 & -3 \end{bmatrix}$$

Would you proceed with using this $\boldsymbol{K}$ to solve the problem? Why or why not?

○ Yes, because $\boldsymbol{K}$ is symmetric.
○ No, because $\boldsymbol{K}$ should be of shape $m \times m$.
● No, because $\boldsymbol{K}$ is not positive semidefinite.
○ No, because $\boldsymbol{K}$ has not been scaled by the variance of the noise.

(b) (5 points) Suppose that a logistic regression model has parameters $w' = (w, b)$. How would the behavior of the model change if parameters were rescaled to $w' \mapsto 2w'$?

○ The output of the model could now take on values within [0,2], reflecting increased classification confidence for the positive class
● The inflection point of the logistic function would have twice the slope
○ The inflection point of the logistic function would now occur at $\langle w', x \rangle = -1$
○ None of the above

(c) (5 points) Suppose there exists a linearly separable dataset D of n samples, where each sample $x_i \in \mathbb{R}^{\mathrm{d}}$ and class label $y_i \in \{\pm 1\}$, and that all samples x_i lie within a d-ball of radius R, which is centered at the origin. Which of the following pre-processing steps would reduce the theoretical error bound on the Perceptron algorithm?

● Applying a feature map of the form $\phi(x) = x(1 - 0.2x)$
○ Applying a feature map of the form $\phi(x) = x^3$
○ Applying a dimensionality reduction technique to decrease d
○ Subsampling the observations to learn a classifier with fewer data points n

(d) (5 points) Robust regression concerns techniques to reduce sensitivity to outliers. A popular loss function is Tukey's biweight loss function

$$l(\gamma) = \begin{cases} \frac{c^2}{6}\left(1 - \left[1 - \left(\frac{\gamma}{c}\right)^2\right]^3\right) & \text{if } |\gamma| \le c, \\ \frac{c^2}{6} & \text{otherwise,} \end{cases}$$

where γ is the residual associated with one observation, and c is a scalar hyperparameter. Which of the following approaches are viable to find the optimal parameters?

○ Deriving a closed-form expression for the optimal parameters

○ Stochastic gradient descent

○ Gradient descent

● Both stochastic gradient descent and gradient descent are appropriate

(e) (5 points) Suppose a linearly separable dataset has 3 samples in the positive class and 4 samples in the negative class. How many variables must one solve for in the dual forms of the hard-margin SVM, and the soft-margin SVM problem, respectively?

○ 7, 14

○ 49, 49

● 7, 7

○ 4, 4

(f) (5 points) What is the primary purpose of using kernel functions in SVMs?

● To allow the model to learn non-linear decision boundaries.

○ To reduce the dimensionality of the data.

○ To regularize the margin.

○ To perform clustering of data points.

(g) (5 points) The posterior distribution in Bayesian learning is:

○ The likelihood of the observed data given the parameters.

● The probability distribution over parameters after observing the data.

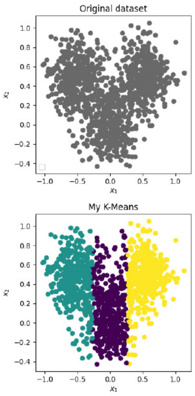
○ The prior distribution before seeing the data.

○ The joint probability of the data and model.

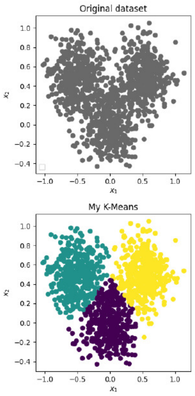
(h) (5 points) Consider the following “buggy” implementation of Lloyd’s algorithm for K-Means clustering:

```
def init_centroids(data_xs, k):
    np.random.shuffle(data_xs)
    centroids = data_xs[:k,:]
    return centroids
def compute_dists(data_xs, centroids):
    dists = np.zeros((data_xs.shape[0], len(centroids)))
    for c, centroid in enumerate(centroids):
        dists[:, c] = np.abs(data_xs[:, 0] - centroid[0])
    return dists
def update_centroids(data_xs, labels, centroids):
    new_centroids = np.zeros((len(centroids), data_xs.shape[1]))
    for i in range(len(centroids)):
        cluster_points = data_xs[labels == i]
        if len(cluster_points) > 0:
            new_centroids[i] = cluster_points.mean(axis=0)
    return new_centroids
def kmeans(data_xs, k, max_iter=100):
    centroids = init_centroids(data_xs, k)
    for i in range(max_iter):
        dists = compute_dists(data_xs, centroids)
        labels = np.argmin(dists, axis=1)
        new_centroids = update_centroids(data_xs, labels, centroids)
        if np.all(centroids == new_centroids):
            break
        centroids = new_centroids
    return centroids, labels
```

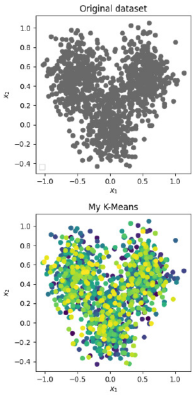
Which of the following clustering outputs would you expect for this code, for a dataset with 2 features, and hyperparameter $k = 3$?



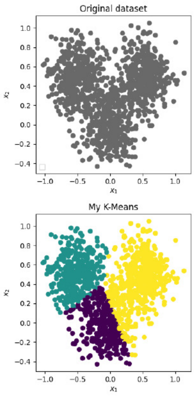
●



○



○



○

Part II: Long-Answer

2. *Linear Regression.* In pharmacology, researchers often study how different dosages of a drug affect patient outcomes, such as symptom relief or the presence of side effects. Typically, the response to a drug increases monotonically, but nonlinearly with the dose:

   At low doses, the drug has little to no effect.

   At moderate doses, the effect increases rapidly.

   At high doses, the effect plateaus as the drug has reached its maximum efficacy.

   (a) (2 points) Propose and write a feature map $\phi : \mathbb{R}^d \mapsto \mathbb{R}^m$, where m is any positive integer, to model the relationships between drug dose and effect in a linear model. Provide a brief justification.

The relationship described b/w dose + effect is similar to that b/w the input and output of a sigmoid function. We can use a set of M sigmoidal basis functions for each drug:

$$\{\varphi_j(x)\} = \left\{\frac{1}{1+e^{-(x-\mu_j)}}\right\}, \quad j = 1, 2, \ldots M.$$

+1: sigmoidal or other feature map w/ nonlinearity

+1: justification/motivation.

   (b) (4 points) Suppose you are part of a team conducting a small clinical trial to study how two drugs interact. These drugs are known to have antagonistic effects, meaning they trigger opposing responses in the body. Your immediate goal is to estimate the parameters of a linear regression model. Identify the approaches you could take to achieve this stated goal, and list the strengths and limitations of each in this particular context.

Some possible approaches are ridge regression, which has the benefit of being able to alleviate overfitting, which is a risk for this small dataset, but wouldn't allow us to incorporate prior knowledge about the drugs' effects. MAP can provide this robustness and incorporate prior knowledge, but introduces bias.

+1 each of RIDGE, MAP, BLR considered.

+1 each. strength/limitation identified.

(Max. 2 marks if considered optimization strategies only)

(c) (4 points) Suppose the team decides to use Maximum A Posteriori estimation. Define and justify a prior distribution for the weights, using the information provided in part (b), and state any assumptions you would make in order to compute the posterior distribution.

We have no reason to believe the weights should take on particular values, so we can specify (assuming M=1)

$$\vec{\mu} = [0, 0]$$

We know the drugs have antagonistic effects, so they should be inversely correlated:

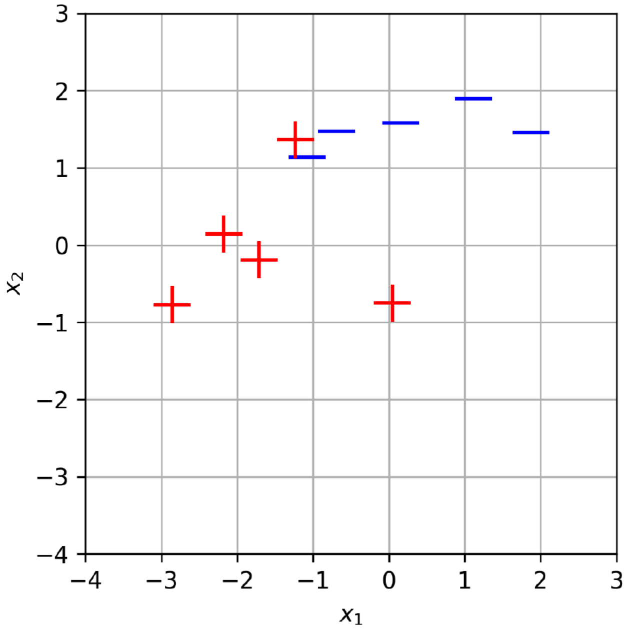
$$\Sigma = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

+2 : Mutivar. gaussian dist for prior

+1 : correct $\vec{\mu}$

+1 : correct Σ

3. *Support Vector Machines.* Consider the following dataset, with the goal of binary classification.



(a) (2 points) Write the corresponding optimization problem that one must solve to find the optimal separating hyperplane. Define all the terms used in the expression.

$$\max_{0 \leq \lambda \leq C} \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{K} \lambda_i \lambda_j y_i y_j \langle x_i , x_j \rangle$$

subject to $\sum_{i=1}^{n} \lambda_i y_i = 0$

where λ_i are the Lagrange multipliers,
y_i are the training data labels, and
x_i are the feature vectors.

+1: SVM objective written (primal or dual is ok)
(−0.5 if anything wrong (ommitted constraints etc)

+1: Variables defined.

(b) (4 points) The training data and output of the optimizer are given in the table below. Compute the parameters of the decision boundary. For expediency, you may use a single observation to compute b.

x_1	x_2	y	λ
-2.86	-0.77	1.	0
0.05	-0.75	1.	107.46
-1.71	-0.19	1.	0
-2.18	0.14	1.	0
-1.23	1.36	1.	888.74
-0.69	1.47	-1.	0
1.12	1.89	-1.	0
-1.08	1.14	-1.	996.21
0.16	1.58	-1.	0
1.87	1.46	-1.	0

② $b = 1 - w^T x$

$$= 1 - [-11.83 \ \ -7.9] \begin{bmatrix} 0.05 \\ -0.75 \end{bmatrix}$$

$$= 1 - (-0.59 + 5.9)$$

$$= 1 - 5.31 = -4.31$$

① $w = \sum \lambda_i y_i x_i$

$$= (107.46)(1) \begin{bmatrix} 0.05 \\ -0.75 \end{bmatrix} = \begin{bmatrix} 5.373 \\ -80.6 \end{bmatrix} = \begin{bmatrix} -11.827 \\ -7.6 \end{bmatrix}$$

$$+ (888.74)(1) \begin{bmatrix} -1.23 \\ 1.36 \end{bmatrix} \quad + \begin{bmatrix} -1093.1 \\ 1208.68 \end{bmatrix}$$

$$+ (996.21)(-1) \begin{bmatrix} -1.08 \\ 1.14 \end{bmatrix} \quad + \begin{bmatrix} 1075.9 \\ -1135.7 \end{bmatrix}$$

$$\boxed{\begin{array}{l} w = [-11.827 \ \ -7.6]^T \\ b = -4.31 \end{array}}$$

+1: formula for w stated

+1: formula for b. stated.

+1: Steps shown.

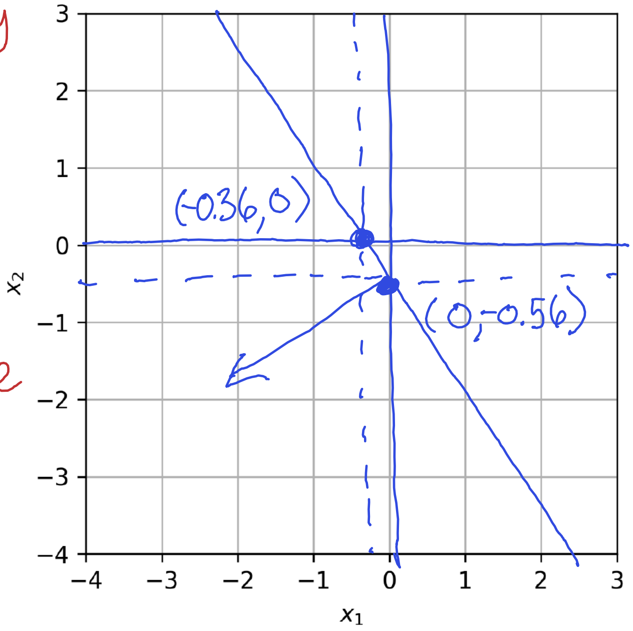
+1: correct answers (or close) (0.5 each of w, b).

(c) (2 points) Sketch the boundary found in part (b) using the space in the graph below. Indicate the direction of the surface normal.

+0.5: A boundary drawn.

+0.5: Surface normal indicated

+1: Some quantitative process:



x_1-int $= \frac{-b}{w_1}$

$\frac{-(-4.3)}{-11.8} = -0.36$

x_2-int $= \frac{-b}{w_2}$

$\frac{-(-4.3)}{-7.6} = -0.56$

(d) (1 point) Suppose that in the training dataset, the point at (-2.86,-0.77) was changed to (-2.86,-1.77). Qualitatively describe the effect on the decision boundary.

The decision boundary would not change because this is not a support vector as λ is zero.

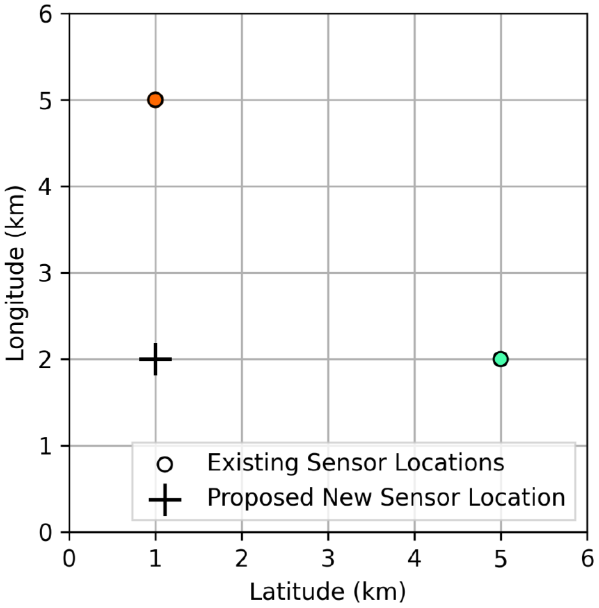
+0.5: correct answer, +0.5: correct justification

(e) (1 point) What can you deduce about the value of any hyperparameter(s) selected to control the optimization process?

The hyperparameter C in the soft-margin SVM must be at least 996, because λ is constrained to be $\leq C$.

+0.5: reference to C; +0.5: ref to λ values

4. *Gaussian Processes.* A group of citizen scientists has been monitoring air quality in a city using a small number of sensors. They have readings from two sensors installed at fixed locations corresponding to average air quality over the past two weeks (Existing Sensor Locations in the figure below). Using these readings, the scientists have proposed to use a Gaussian Process model to predict the air quality at unsampled locations and quantify their confidence in those predictions. They also have the opportunity to install one additional sensor, and one particular location is under consideration (Proposed New Sensor Location in the figure below).



(a) (2 points) Defend the choice of a Gaussian Process model for this application, and recommend and justify a kernel function for the model.

Gaussian processes are appropriate for this problem as we want to interpolate b/w missing data points, have sparse data, and cannot specify the form of the true underlying function. A Gaussian Kernel is appropriate as air quality is likely to be locally smooth.

+1 At least one justification.

+1 Valid Kernel+reasoning.

(b) (8 points) Compute the predictive mean for the candidate sensor at location (1.0, 2.0), using the air quality readings (AQI) from the existing sensors in this table:

	Sensor ID	Latitude	Longitude	AQI
X_1	001	1.0	5.0	100
X_2	002	5.0	2.0	10
X_t.		1.0	2.0.	

Assume a Gaussian prior over the weights with precision $\alpha = 1$, and that you have noise-free observations (i.e. $\beta \to \infty$). Use the stationary kernel

$$k(x, x') = \max(0, 6 - \|x - x'\|),$$

where $\|v\|$ denotes the L2 norm for some $v \in \mathbb{R}^d$. The following formulae may be useful:

$$\mu(x_{N+1}) = \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{t} \qquad \mathbf{C}_{N+1} = \begin{pmatrix} \mathbf{C}_N & \mathbf{k} \\ \mathbf{k}^\top & c \end{pmatrix} \qquad \mathbf{C}_{ij} = \alpha^{-1} k(x_i, x_j) + \beta^{-1} \delta_{ij}$$

$K(x_1, x_2) = 6 - 5 = 1$

$K(x_1, x_t) = 6 - 3 = 3.$

$K(x_2, x_t) = 6 - 4 = 2.$

$C_N = \begin{bmatrix} 6 & 1 \\ 1 & 6 \end{bmatrix} \qquad C_N^{-1} = \frac{1}{36-1} \begin{bmatrix} 6 & -1 \\ -1 & 6 \end{bmatrix}$

$k^T = \begin{bmatrix} 3 & 2 \end{bmatrix} \qquad \mu = \begin{bmatrix} 3 & 2 \end{bmatrix} \frac{1}{35} \begin{bmatrix} 6 & -1 \\ -1 & 6 \end{bmatrix} \begin{bmatrix} 100 \\ 10 \end{bmatrix}$

$t = \begin{bmatrix} 100 \\ 10 \end{bmatrix}$

$= \frac{1}{35} \begin{bmatrix} 18\text{-}2 & -3+12 \end{bmatrix} \begin{bmatrix} 100 \\ 10 \end{bmatrix} = \frac{1}{35} \begin{bmatrix} 1600 + 90 \end{bmatrix}$

$= 1690/35$

$= 48.4$

+4 compute pairwise interactions

+1 construct C_N.

+1 construct K^T

+1 construct t.

+1 correct answer (or close)

Part III: Short Answer

+2 correct answer
+3 reasonable justification.

5. For each question, answer whether it is **True** or **False**, and provide a brief (1 sentence) justification for your answer.

(a) (5 points) Decision trees are invariant to scaling of the input features.

Multiple answers accepted, provided answer and justification were coherent. (see next page)

(b) (5 points) The error bound of the Perceptron algorithm indicates the precise number of errors that will be made during training.

false. It is only an upper bound.

(c) (5 points) The kernel trick allows us to use any similarity metric in high-dimensional space.

True. If it wasnt a similarity metric, it would not be a valid Kernel.

(d) (5 points) A function of more than one variable is convex if and only if its Hessian is everywhere positive definite.

False. A function that is everywhere twice differentiable is convex if its Hessian is everywhere positive Semidefinite.

(e) (5 points) Least-squares regression assumes that the dependent variable is normally distributed.

False. Least-squares regression assumes the errors are normally distributed.

(f) (5 points) Logistic regression cannot handle multi-class classification problems.

False. Logistic regression can be generalized to the multiclass case using softmax.

This page is intentionally left blank to accommodate work that wouldn't fit elsewhere and/or scratch work.

Answers accepted to Q5a.

False. The thresholds would be different.

True. Decision trees can split on any feature + threshold so if the features are scaled uniformly, the resulting tree would behave the same.

This page is intentionally left blank to accommodate work that wouldn't fit elsewhere and/or scratch work.

Q1a. A positive semidefinite matrix A has non-negative eigenvalues. The eigenvalues may be determined by solving

$$|A - \lambda I| = 0,$$ where $|\cdot|$ is the determinant

$$K - \lambda I = \begin{bmatrix} 1 & 2 \\ 2 & -3 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = \begin{bmatrix} 1-\lambda & 2 \\ 2 & -3-\lambda \end{bmatrix}$$

$$|K - \lambda I| = (1-\lambda)(-3-\lambda) - (2)(2)$$
$$= -3 - \lambda + 3\lambda + \lambda^2 - 4 = \lambda^2 + 2\lambda - 7$$
$$= (\lambda + a)(\lambda + b),$$ where

$$ab = -7, \text{ and } a + b = 2$$

To satisfy both conditions, either a or b must be negative, so one of the eigenvalues must be negative, meaning that K cannot be PSD.

So, we should not proceed with using K to solve the problem, because K is not positive semidefinite.

This page is intentionally left blank to accommodate work that wouldn't fit elsewhere and/or scratch work.

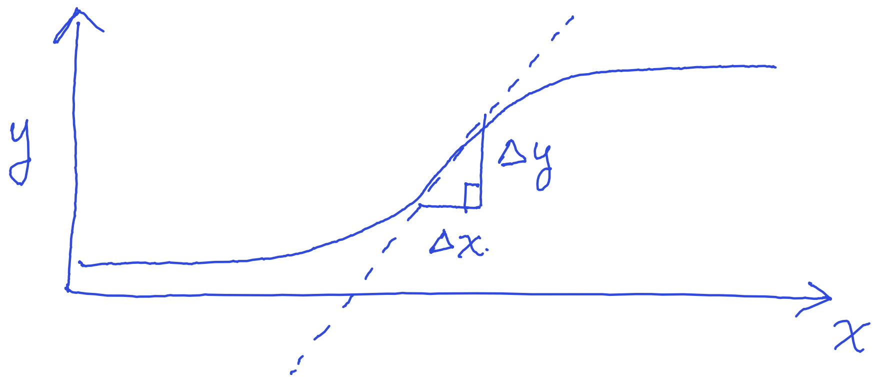
Q1b The logistic function is given by

$$y = \frac{1}{1+e^{-(wx+b)}}$$

→ if $t = wx+b$ is the preactivation, y is bounded to within the range of $[0,1)$ $\forall t$

→ only changing b would change the position of the inflection point.

→ at the inflection point, $\Delta y \propto w \Delta x$. therefore $w \rightarrow 2w \Rightarrow \Delta y \rightarrow 2\Delta y$



This page is intentionally left blank to accommodate work that wouldn't fit elsewhere and/or scratch work.

Q1c. The Perceptron error bound is given by

$$K \leq \left(\frac{R}{\gamma}\right)^2.$$

This theoretical result is independent of the number of samples or dimensionality of the data.

To reduce K, we need to reduce the ratio R/γ.

We were given that the data lies within a d-Ball of radius $R=1$, centered at the origin and that feature maps are applied independently to each input feature.

$\varphi(x) = x^3$: $R' = R$ but $\gamma' < \gamma$. this would increase the ratio R/γ.

$\varphi(x) = x(1-0.2x)$: $\gamma' < \gamma$ but $R' \lll R$, thus providing an overall reduction in R/γ.

Q1d We can split the dataset into two subsets, one for which $\gamma \leq c$, and one for which $\gamma > c$.

We can then compute then compute the respective gradients either across the whole dataset (GD) or subsample minibatches. (SGD).

Q1e. The dual forms of the SVM problems — either hard-margin or softmargin — have one variable per data point. As there are 7 samples total, the answer is 7, 7.

Q1f. It is generally true that we use a kernel function in SVM if we needed a nonlinear decision boundary. often, kernel functions enable more expressive classifier behavior as if it were operating in a higher-dimensional feature space, but this is not necessarily true.

Q1g The posterior distribution is

$$p(w|x) \propto p(x|w)p(w)$$

where $p(w)$ is the prior over parameters, $p(x|w)$ is the likelihood of the data, and $p(w|x)$ is the probability of the parameters after observing the data.

Q1h. The distance metric in Line 8 reveals that only the absolute difference along the first dimension is considered. This would explain clustering output of the leftmost plot.