Randall Rogers
DSC 255: Machine Learning Fundamentals
Homework 3
Spring 2025

## Question 1

Gaussian parameters. Each of the following scenarios describes a joint distribution $(x, y)$. In each case, give the parameters of the (unique) bivariate Gaussian that satisfies these properties.

(a) x has mean 2 and standard deviation 1, y has mean 2 and standard deviation 0.5, and the correlation between x and y is -0.5.

(b) x has mean 1 and standard deviation 1, and y is equal to 2.

**Solution (a)**

---

For a bivariate Gaussian distribution, we need to specify the following parameters:

$$\boldsymbol{\mu} = [\mu_x, \mu_y]^T$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$$

- Mean vector: $\boldsymbol{\mu}$
- Covariance matrix: $\boldsymbol{\Sigma}$

Let $\mu_x = 2$, $\mu_y = 2$, $\sigma_x = 1$, $\sigma_y = 0.5$, and $\rho = -0.5$

Solve for mean vector $\boldsymbol{\mu}$:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_x & \mu_y \end{bmatrix}^T = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

Solve for covariance matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} = \begin{bmatrix} 1^2 & (-0.5)(1)(0.5) \\ (-0.5)(1)(0.5) & 0.5^2 \end{bmatrix} = \begin{bmatrix} 1 & -0.25 \\ -0.25 & 0.25 \end{bmatrix}$$

$\therefore$ the bivariate Gaussian has parameters:

$$\mathcal{N}\left( \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & -0.25 \\ -0.25 & 0.25 \end{bmatrix} \right)$$

---

**Solution (b)**

---

For a bivariate Gaussian distribution, we need to specify the following parameters:

$$\boldsymbol{\mu} = [\mu_x, \mu_y]^T$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$$

- Mean vector: $\boldsymbol{\mu}$
- Covariance matrix: $\boldsymbol{\Sigma}$

Let $\mu_x = 1$, $\mu_y = 2$, $\sigma_x = 1$, $\sigma_y = 0$

Solve for mean vector $\boldsymbol{\mu}$:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_x & \mu_y \end{bmatrix}^T = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Solve for covariance matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} = \begin{bmatrix} 1^2 & 0 \\ 0 & 0^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

$\therefore$ the bivariate Gaussian has parameters:

$$\mathcal{N}\left( \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \right)$$

---

---

## Question 2

A generative approach is used for a binary classifcation problem (with classes +, -) and it turns out that the resulting classifier predicts + at all points x in the input space. Why might this be?

**Solution**

---

**Step 1:** Recall the Bayesian decision rule for generative classifiers.

In a generative approach, we model the joint distribution $p(x, y)$ by learning the class-conditional densities $p(x|y)$ and prior probabilities $p(y)$. The classification decision is made using Bayes' rule:

$$\hat{y} = \arg \max_{y \in \{+,-\}} p(y|x) = \arg \max_{y \in \{+,-\}} \frac{p(x|y)p(y)}{p(x)}$$

Since $p(x)$ is constant for both classes, the decision rule simplifies to:

$$\hat{y} = \arg \max_{y \in \{+,-\}} p(x|y)p(y)$$

**Step 2:** Analyze the decision boundary.

For binary classification, we predict the positive class (+) when:

$$p(x|+)p(+) > p(x|-)p(-)$$

Equivalently, taking the logarithm (which preserves the inequality since log is monotonically increasing):

$$\log p(x|+) + \log p(+) > \log p(x|-) + \log p(-)$$

**Step 3:** Identify possible reasons for always predicting the positive class.

There are several reasons why the inequality $p(x|+)p(+) > p(x|-)p(-)$ might hold for all $x$:

1. **Highly imbalanced prior probabilities:** If $p(+) \gg p(-)$, the classifier might always predict the positive class because the prior term dominates the decision, regardless of the likelihood term. This occurs when the training data contains many more positive examples than negative ones.
2. **Poor estimation of class-conditional densities:** If $p(x|+)$ is consistently overestimated or $p(x|-)$ is consistently underestimated across the input space, the classifier will favor the positive class.
3. **Model misspecification:** If the assumed parametric form of $p(x|y)$ (e.g., Gaussian) does not match the true distribution of the data, the estimated densities may lead to systematic classification errors.
4. **Feature insufficiency:** If the selected features do not provide discriminative information to separate the classes, the model might default to the more common class.
5. **Numerical issues:** In high-dimensional spaces, numerical underflow when computing likelihoods can lead to computational issues that affect the decision boundary.

**Step 4:** Mathematical illustration with Gaussian class-conditionals.

For a concrete example, consider a generative classifier with Gaussian class-conditionals:

$$p(x|+) \sim \mathcal{N}(\mu_+, \Sigma_+)$$
$$p(x|-) \sim \mathcal{N}(\mu_-, \Sigma_-)$$

The decision boundary is determined by:

$$\log p(x|+) + \log p(+) = \log p(x|-) + \log p(-)$$

Expanding the Gaussian log-likelihoods:

$$-\frac{1}{2}(x-\mu_+)^T\Sigma_+^{-1}(x-\mu_+) - \frac{1}{2}\log|\Sigma_+| - \frac{d}{2}\log(2\pi) + \log p(+) =$$
$$-\frac{1}{2}(x-\mu_-)^T\Sigma_-^{-1}(x-\mu_-) - \frac{1}{2}\log|\Sigma_-| - \frac{d}{2}\log(2\pi) + \log p(-)$$

If this equation has no solution for any $x$ (i.e., one side is always greater), the classifier will always predict the same class. This can happen if:

- The means $\mu_+$ and $\mu_-$ are very close, but $p(+) \gg p(-)$
- The covariance matrices $\Sigma_+$ and $\Sigma_-$ are poorly estimated
- The dimensionality $d$ is high relative to the sample size, leading to unreliable parameter estimates

**Conclusion:** A generative classifier that always predicts the positive class indicates an issue with either the data distribution (class imbalance), the model specification, or the parameter estimation process. In practice, this suggests the need to:

- Balance the training data or adjust class priors
- Try different parametric forms for the class-conditional densities
- Apply regularization to improve parameter estimation
- Consider dimensionality reduction if working in high-dimensional spaces
- Validate the model's assumptions about the data distribution

---

---

## Question 3

Winery classification. For the winery example from lecture, the densities obtained are reproduced here:

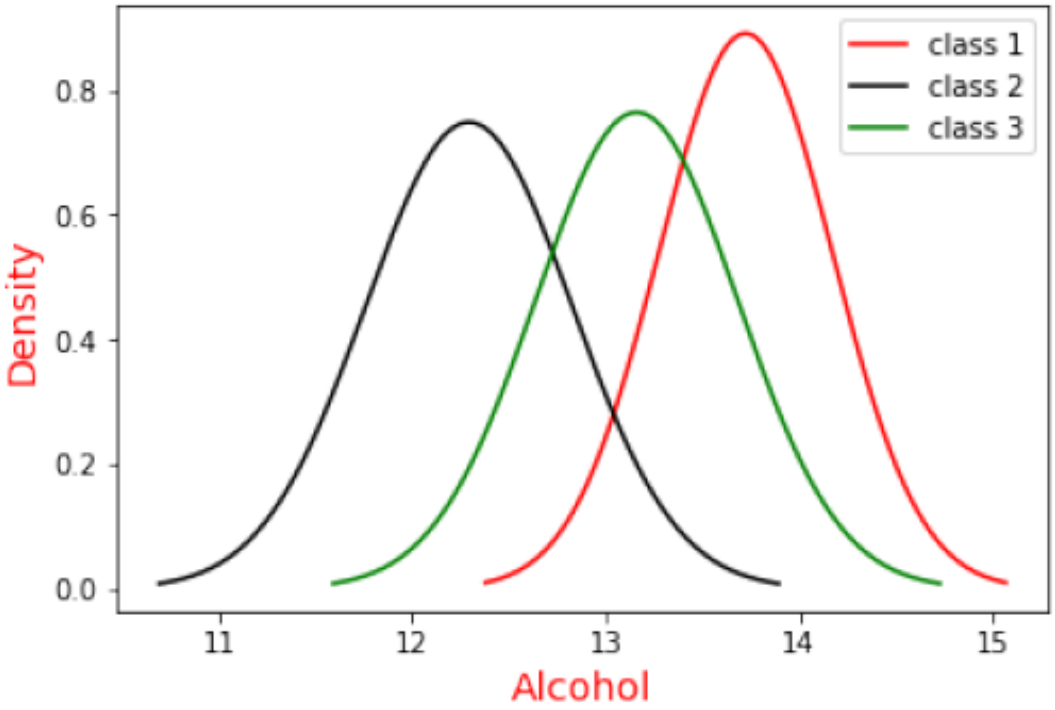



Figure 1: Winery classification densities

The class probabilities are $\pi_1 = 0.33$; $\pi_2 = 0.39$; $\pi_3 = 0.28$. What labels would be assigned to the following points?

(a) 12.0

(b) 12.5

(c) 13.0

(d) 13.5

(e) 14.0

**Solution (a)**

---

From Bayes Theorem we have:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

*or*

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

The class probabilities are given as: $\pi_1 = 0.33$; $\pi_2 = 0.39$; $\pi_3 = 0.28$

From *Figure 1*, we can see that the density for each class 1 at $x = 12.0$ is: $P(x = 12.0|Class_1) = 0.4$; for class 2 is $P(x = 12.0|Class_2) = 0.05$; and for class 3 is $P(x = 12.0|Class_3) = 0.0025$.

Apply Bayes Theorem approimation to solve for $P(Class_i|x = 12.0)$

$$P(Class_i|x) \propto P(x|Class_i)P(Class_i)$$

Find $P(Class_i|x = 12.0)$ for each class:

$P(Class_1|x = 12.0) \propto P(x = 12.0|Class_1)P(Class_1) = 0.4 \times 0.33 = 0.132$

$P(Class_2|x = 12.0) \propto P(x = 12.0|Class_2)P(Class_2) = 0.05 \times 0.39 = 0.0195$

$P(Class_3|x = 12.0) \propto P(x = 12.0|Class_3)P(Class_3) = 0.0025 \times 0.28 = 0.0007$

$\therefore$ the label $Class_1$ would be assigned at x=12.0

---

**Solution (b)**

---

From Bayes Theorem we have:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

$$or$$

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

The class probabilities are given as: $\pi_1 = 0.33$; $\pi_2 = 0.39$; $\pi_3 = 0.28$

From *Figure 1*, we can see that the density for each class 1 at $x = 12.5$ is: $P(x = 12.5|Class_1) = 0.6$; for class 2 is $P(x = 12.5|Class_2) = 0.3$; and for class 3 is $P(x = 12.5|Class_3) = 0.05$.

Apply Bayes Theorem approimation to solve for $P(Class_i|x = 12.5)$

$$P(Class_i|x) \propto P(x|Class_i)P(Class_i)$$

Find $P(Class_i|x = 12.5)$ for each class:

$P(Class_1|x = 12.5) \propto P(x = 12.5|Class_1)P(Class_1) = 0.6 \times 0.33 = 0.198$

$P(Class_2|x = 12.5) \propto P(x = 12.5|Class_2)P(Class_2) = 0.3 \times 0.39 = 0.117$

$P(Class_3|x = 12.5) \propto P(x = 12.5|Class_3)P(Class_3) = 0.05 \times 0.28 = 0.014$

$\therefore$ the label $Class_1$ would be assigned at x=12.5

---

**Solution (c)**

---

From Bayes Theorem we have:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

$$or$$

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

The class probabilities are given as: $\pi_1 = 0.33$; $\pi_2 = 0.39$; $\pi_3 = 0.28$

From *Figure 1*, we can see that the density for each class 1 at $x = 13.0$ is: $P(x = 13.0|Class_1) = 0.3$; for class 2 is $P(x = 13.0|Class_2) = 0.6$; and for class 3 is $P(x = 13.0|Class_3) = 0.2$.

Apply Bayes Theorem approimation to solve for $P(Class_i|x = 13.0)$

Find $P(Class_i|x = 13.0)$ for each class:

$$P(Class_i|x) \propto P(x|Class_i)P(Class_i)$$

$P(Class_1|x = 13.0) \propto P(x = 13.0|Class_1)P(Class_1) = 0.3 \times 0.33 = 0.099$

$P(Class_2|x = 13.0) \propto P(x = 13.0|Class_2)P(Class_2) = 0.6 \times 0.39 = 0.234$

$P(Class_3|x = 13.0) \propto P(x = 13.0|Class_3)P(Class_3) = 0.2 \times 0.28 = 0.056$

$\therefore$ the label $Class_2$ would be assigned at x=13.0

---

**Solution (d)**

---

From Bayes Theorem we have:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

$$or$$

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

The class probabilities are given as: $\pi_1 = 0.33$; $\pi_2 = 0.39$; $\pi_3 = 0.28$

From *Figure 1*, we can see that the density for each class 1 at $x = 13.5$ is: $P(x = 13.5|Class_1) = 0.1$; for class 2 is $P(x = 13.5|Class_2) = 0.7$; and for class 3 is $P(x = 13.5|Class_3) = 0.4$.

Apply Bayes Theorem approimation to solve for $P(Class_i|x = 13.5)$

Find $P(Class_i|x = 13.5)$ for each class:

$$P(Class_i|x) \propto P(x|Class_i)P(Class_i)$$

$P(Class_1|x = 13.5) \propto P(x = 13.5|Class_1)P(Class_1) = 0.1 \times 0.33 = 0.033$

$P(Class_2|x = 13.5) \propto P(x = 13.5|Class_2)P(Class_2) = 0.7 \times 0.39 = 0.273$

$P(Class_3|x = 13.5) \propto P(x = 13.5|Class_3)P(Class_3) = 0.4 \times 0.28 = 0.112$

$\therefore$ the label $Class_2$ would be assigned at x=13.5

---

**Solution (e)**

---

From Bayes Theorem we have:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

$$or$$

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

The class probabilities are given as: $\pi_1 = 0.33$; $\pi_2 = 0.39$; $\pi_3 = 0.28$

From *Figure 1*, we can see that the density for each class 1 at $x = 14.0$ is: $P(x = 14.0|Class_1) = 0.05$; for class 2 is $P(x = 14.0|Class_2) = 0.2$; and for class 3 is $P(x = 14.0|Class_3) = 0.8$.

Apply Bayes Theorem approimation to solve for $P(Class_i|x = 14.0)$

Find $P(Class_i|x = 14.0)$ for each class:

$$P(Class_i|x) \propto P(x|Class_i)P(Class_i)$$

$P(Class_1|x = 14.0) \propto P(x = 14.0|Class_1)P(Class_1) = 0.05 \times 0.33 = 0.0165$

$P(Class_2|x = 14.0) \propto P(x = 14.0|Class_2)P(Class_2) = 0.2 \times 0.39 = 0.078$

$P(Class_3|x = 14.0) \propto P(x = 14.0|Class_3)P(Class_3) = 0.8 \times 0.28 = 0.224$

$\therefore$ the label $Class_3$ would be assigned at x=14.0

---

---

## Question 4

Gaussian contours. Roughly sketch the shapes of the following Gaussians $N(\mu; \Sigma)$. You only need to show a representative contour line which is qualitatively accurate (has approximately the right orientation, for instance).

(a) $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$; $\Sigma = \begin{bmatrix} 9 & 0 \\ 0 & 1 \end{bmatrix}$

(b) $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$; $\Sigma = \begin{bmatrix} 1 & -0.75 \\ -0.75 & 1 \end{bmatrix}$

**Solution (a)**

---

Analyze $\mu$:

- Since $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, the center of the Gaussian is at the origin.

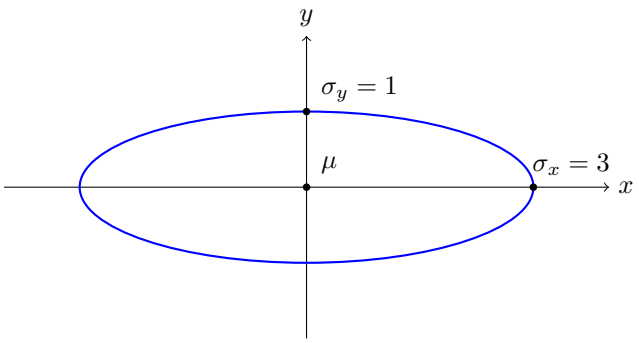
Analyze $\Sigma$:

- The covariance matrix $\Sigma = \begin{bmatrix} 9 & 0 \\ 0 & 1 \end{bmatrix}$ indicates that the variance in the x-direction is 9 and in the y-direction is 1. This means that the Gaussian will be elongated along the x-axis.

Look at the standard deviation to determine the shape, since $\Sigma$ is a diagonal matrix:

- The standard deviation in the x-direction is $\sigma_x = \sqrt{9} = 3$
- The standard deviation in the y-direction is $\sigma_y = \sqrt{1} = 1$

Sketch the Gaussian




---

**Solution (b)**

---

Analyze $\mu$:

- Since $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, the center of the Gaussian is at the origin.

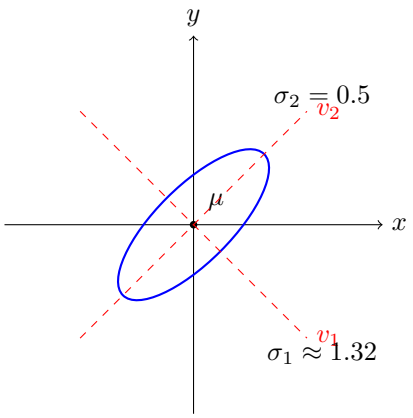
Analyze $\Sigma$:

- The covariance matrix $\Sigma = \begin{bmatrix} 1 & -0.75 \\ -0.75 & 1 \end{bmatrix}$ has non-zero off-diagonal elements.
- This indicates a correlation between variables.
- The negative correlation $(-0.75)$ means that as one variable increases, the other tends to decrease.
- Look at the eigenvalues and eigenvectors to determine the shape since we don't have a diagonal matrix in this case.

Find eigenvalues and eigenvectors:

- The eigenvalues of $\Sigma$ are $\lambda_1 = 1 + 0.75 = 1.75$ and $\lambda_2 = 1 - 0.75 = 0.25$.
- The corresponding eigenvectors are $v_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ and $v_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.
- The standard deviations along the principal axes are $\sigma_1 = \sqrt{1.75} \approx 1.32$ and $\sigma_2 = \sqrt{0.25} = 0.5$.

Sketch the Gaussian



## Question 5

Find all unit vectors in $\mathbb{R}^2$ that are orthogonal to $\begin{bmatrix}1\\1\end{bmatrix}$.

**Solution**

---

A unit vector in $\mathbb{R}^2$ is a vector of the form:

$$\begin{bmatrix}x\\y\end{bmatrix}$$

To be orthogonal to $\begin{bmatrix}1\\1\end{bmatrix}$, the dot product must equal zero:

$$\begin{bmatrix}1\\1\end{bmatrix} \cdot \begin{bmatrix}x\\y\end{bmatrix} = 0$$

This gives the equation:

$$x + y = 0$$

This means that $y = -x$.
Now, we need to find the unit vectors. A unit vector has a magnitude of 1:

$$\sqrt{x^2 + y^2} = 1$$

Substituting $y = -x$ into the equation:

$$\sqrt{x^2 + (-x)^2} = 1$$
$$\sqrt{2x^2} = 1$$
$$\sqrt{2}|x| = 1$$
$$|x| = \frac{1}{\sqrt{2}}$$

This gives two solutions for $x$:

$$x = \frac{1}{\sqrt{2}} \quad \text{or} \quad x = -\frac{1}{\sqrt{2}}$$

Substituting back to find $y$:

$$y = -\frac{1}{\sqrt{2}} \quad \text{or} \quad y = \frac{1}{\sqrt{2}}$$

$\therefore$ the unit vectors orthogonal to $\begin{bmatrix}1\\1\end{bmatrix}$ are:

$$\begin{bmatrix}\frac{1}{\sqrt{2}}\\-\frac{1}{\sqrt{2}}\end{bmatrix} \quad \text{and} \quad \begin{bmatrix}-\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}\end{bmatrix}$$

---

## Question 6

How would you describe the set of all points $x \in \mathbb{R}^d$ with $x \cdot x = 25$
**Solution**

---

The equation $x \cdot x = 25$ describes a sphere in $\mathbb{R}^d$ with radius 5.
The set of all points $x$ that satisfy this equation is the surface of a sphere centered at the origin with radius 5.
In $d$-dimensional space, the equation can be written as:

$$x_1^2 + x_2^2 + \ldots + x_d^2 = 25$$

This represents a $d$-dimensional sphere (or hypersphere) with radius 5.
The set of all points $x$ that satisfy this equation is the surface of a sphere in $\mathbb{R}^d$.
The surface of the sphere is a $(d-1)$-dimensional manifold embedded in $\mathbb{R}^d$.
In summary, the set of all points $x \in \mathbb{R}^d$ with $x \cdot x = 25$ is a sphere of radius 5 centered at the origin in $d$-dimensional space.
$\therefore$ the set of all points $x \in \mathbb{R}^d$ with $x \cdot x = 25$ is a sphere of radius 5 centered at the origin in $d$-dimensional space.

---

---

## Question 7

Consider the function $f(x) = 2x_1 - x_2 + 6x_3$ can be written as $w \cdot x$ for $x \in \mathbb{R}^3$. What is $w$
**Solution**

---

The function $f(x) = 2x_1 - x_2 + 6x_3$ can be expressed in the form of a dot product:

$$f(x) = w \cdot x$$

Where $w$ is a vector in $\mathbb{R}^3$ and $x$ is a vector in $\mathbb{R}^3$.

We can then express $f(x)$ as the dot product of two matrices:

$$f(x) = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = w_1 \cdot x_1 + w_2 \cdot x_2 + w_3 \cdot x_3$$

It follows that: $w_1 = 2 \;\; w_2 = -1 \;\; w_3 = 6$

$\therefore$ the vector $w$ is:

$$w = \begin{bmatrix} 2 \\ -1 \\ 6 \end{bmatrix}$$

---

---

## Question 8

For a certain pair of matrices $A$, $B$ the product $AB$ has dimension $10 \times 20$. If $A$ has 30 columns, what are the dimensions of $A$ and $B$?

**Solution**

---

Let the dimensions of $A$ be $m \times n$ and the dimensions of $B$ be $n \times p$.

The product $AB$ will have dimensions $m \times p$.

Given that $AB$ has dimensions $10 \times 20$, and $A = m \times 30$:

$$m = 10 \quad \text{and} \quad p = 20$$

Also, we know that $A$ has 30 columns, which means:

$$n = 30$$

$\therefore$ the dimensions of $A$ and $B$ are:

$$A : 10 \times 30$$

$$B : 30 \times 20$$

---

---

## Question 9

We have $n$ data points $x^{(1)}$,...,$x^{(n)} \in \mathbb{R}^d$ and we store them in a matrix $X$, one point per row.

$$X = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(n)} \end{bmatrix}$$

(a) What is the dimension of $X$, in terms of n and d?

(b) What is the dimension of $XX^T$?

(c) What is the $(i, j)$ entry of $X^TX$, simply?

**Solution (a)**

---

The matrix $X$ has $n$ rows and $d$ columns, so the dimension of $X$ is:

$$X \in \mathbb{R}^{n \times d}$$

This means that $X$ has $n$ data points, each with $d$ features.

$\therefore$ the dimension of $X$ is $n \times d$.

---

**Solution (b)**

---

The matrix $XX^T$ is the product of an $n \times d$ matrix and a $d \times n$ matrix.

The resulting matrix will have dimensions $n \times n$.

$\therefore$ the dimension of $XX^T$ is $n \times n$.

---

**Solution (c)**

---

The $(i, j)$ entry of $X^TX$ is the dot product of the $i$-th row of $X$ and the $j$-th column of $X^T$

This is simply the sum of the products of the corresponding elements:

$$(X^TX)_{ij} = \sum_{k=1}^{d} x_k^{(i)} x_k^{(j)}$$

This is the inner product of the $i$-th and $j$-th data points.

$\therefore$ the $(i, j)$ entry of $X^TX$ is the inner product of the $i$-th and $j$-th data points or $(x^{(n)}, x^{(n)})$.

---

---

## Question 10

For $x = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix}$, compute $x^T x$ and $xx^T$.

**Solution**

---

The vector $x$ is:

$$x = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix}$$

To compute $x^T x$:

$$x^T x = \begin{bmatrix} 1 & 3 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix}$$

This is the dot product of $x$ with itself:

$$x^T x = 1^2 + 3^2 + 5^2 = 1 + 9 + 25 = 35$$

To compute $xx^T$:

$$xx^T = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} \begin{bmatrix} 1 & 3 & 5 \end{bmatrix}$$

This is the outer product of $x$ with itself:

$$xx^T = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} \begin{bmatrix} 1 & 3 & 5 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 5 \\ 3 & 9 & 15 \\ 5 & 15 & 25 \end{bmatrix}$$

$\therefore$ the results are:

$$x^T x = 35$$

$$xx^T = \begin{bmatrix} 1 & 3 & 5 \\ 3 & 9 & 15 \\ 5 & 15 & 25 \end{bmatrix}$$

---

---

## Question 11

The quadratic function $f : \mathbb{R}^3 \to \mathbb{R}$ given by:

$$f(x) = 3x_1^2 + 2x_1x_2 - 4x_1x_3 + 6x_3^2$$

can be written in the form $x^T Mx$ for some symmetic matrix $M$. What is $M$?

**Solution**

---

The quadratic function can be expressed in the form:

$$f(x) = x^T Mx$$

Where $x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ and $M$ is a symmetric matrix.

To find the matrix $M$, we can rewrite the quadratic function in matrix form:

$$f(x) = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} 3 & 1 & -2 \\ 1 & 0 & 0 \\ -2 & 0 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

This gives us the matrix $M$:

$$M = \begin{bmatrix} 3 & 1 & -2 \\ 1 & 0 & 0 \\ -2 & 0 & 6 \end{bmatrix}$$

To ensure that $M$ is symmetric, we can check that $M = M^T$:

$$M^T = \begin{bmatrix} 3 & 1 & -2 \\ 1 & 0 & 0 \\ -2 & 0 & 6 \end{bmatrix}$$

This confirms that $M$ is symmetric.

$\therefore$ the matrix $M$ is:

$$M = \begin{bmatrix} 3 & 1 & -2 \\ 1 & 0 & 0 \\ -2 & 0 & 6 \end{bmatrix}$$

---

---

## Question 12

Let $A = diag(1, 2, 3, 4, 5, 6, 7, 8)$:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8 \end{bmatrix}$$

(a) What is $|A|$?

(b) What is $A^{-1}$?

**Solution (a)**

---

The determinant of a diagonal matrix is the product of its diagonal elements:

$$|A| = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 = 8! = 40320$$

$\therefore$ the determinant of $A = 40320$

---

**Solution (b)**

---

The inverse of a diagonal matrix is obtained by taking the reciprocal of each diagonal element:

$$A^{-1} = diag\left(\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \frac{1}{8}\right)$$

$\therefore$ the inverse of $A$ is:

$$A^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{5} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{6} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{7} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{8} \end{bmatrix}$$

---

---

## Programming Exercises

## Question 13

Classifying MNIST digits using generative modeling. In class, we have already encountered the MNIST data set of handwritten digits. In this problem, you will build a classifier for this data, by modeling each class as a multivariate (784-dimensional) Gaussian.

- Download the Jupyter notebook `generative-mnist.ipynb` from the course website. Look over the notebook to see what it is doing, and then run it, one cell at a time. Make sure you understand the form in which the training and test data are stored. Towards the end of the notebook, there is also a helper function that displays digits.
- Split the training set into two pieces a training set of size 50000 (say), and a separate validation set of size 10000.
- Now fit a Gaussian generative model to the training data of 50000 points. Determine the class probabilities: what fraction $\pi_0$ of the training points are digit 0, for instance? Call these values $\pi_0, ..., \pi_9$. Fit a Gaussian to each digit, by finding the mean and the covariance of the corresponding data points. Let the Gaussian for the jth digit be $P_j = N(\mu_j, \Sigma_j)$. Note that $\mu_j$ will be a 784-dimensional vector, and $\Sigma_j$ will be a $784 \times 784$ matrix. Using these two pieces of information, you can classify new images using Bayes' rule: simply pick the digit j for which $\pi_j P_j(x)$ is largest.
- One last step is needed: it is important to smooth the covariance matrices, and the usual way to do this is to add in $cI$, where c is some constant and I is the identity matrix. Use the validation set to help you choose the right value of c: that is, choose the value of c for which the resulting classifier makes the fewest mistakes on the validation set.
- There are some important details of numerical precision that deserve attention. In 784-dimensional space, all probabilities $P_j(x)$ will likely be miniscule, and this can produce all sorts of trouble due to underflow errors. It is better to work with log-probabilities: 1000 is easier to deal with than $e^{-1000}$. This means that you should classify a point by picking the j that maximizes $log\ \pi_j + log\ P_j(x)$. Fortunately, the Python multivariate normal package will directly compute $log\ P_j(x)$ for you.
- To turn in:
  (a) Pseudocode for your training procedure, making it clear how the validation set was created and used.
  (b) Did you use a single value of c for all ten classes, or separate values for each class? What value(s) of c did you get?
  (c) What was the error rate on the MNIST test set?
  (d) Out of the misclassified test digits, pick five at random and display them.

**13 (a)**

---

Pseudocode for training procedure:

1. Input
    - x: training data
    - y: training labels
    - c: smoothing constant for covariance matrices
2. Initialize
    - x_train: 80% of x
    - x_val: 20% of x
    - y_train: 80% of y
    - y_val: 20% of y
    - $\pi_k$: class $k$ frequencies
    - $\mu_k$: class $k$ mean vector
    - $\Sigma_k$: class $k$ covariance matrices
    - qda: Gaussian generative model
3. Iterate
    - For each class $k$ in 0, 1, ..., 9:
        - Compute $\pi_k$ as the fraction of training points in class $k$
        - Compute $\mu_k$ as the mean of training points in class $k$
        - Compute $\Sigma_k$ as the covariance of training points in class $k$
        - Smooth $\Sigma_k$ by adding $cI$ to it
        - Store each $\pi_k$, $\mu_k$, and $\Sigma_k$
4. Evaluate
    - For each point in the validation set(x_val) and each c value tested:
        - Compute average accuracy of the model over all classes
        - Store the accuracy for each c value
        - Choose the c value that gives the best accuracy
        - Store the best c value
    - Compute the final model using the best c value
    - Store each $\pi_k$, $\mu_k$, and $\Sigma_k$ for the final model

**13 (b)**

- A single value of $c = 0.95$ for all ten classes.
- The value of $c$ was chosen based on the validation set accuracy.
- The accuracy for $c = 0.95$ was 0.8742 or 87.42%.

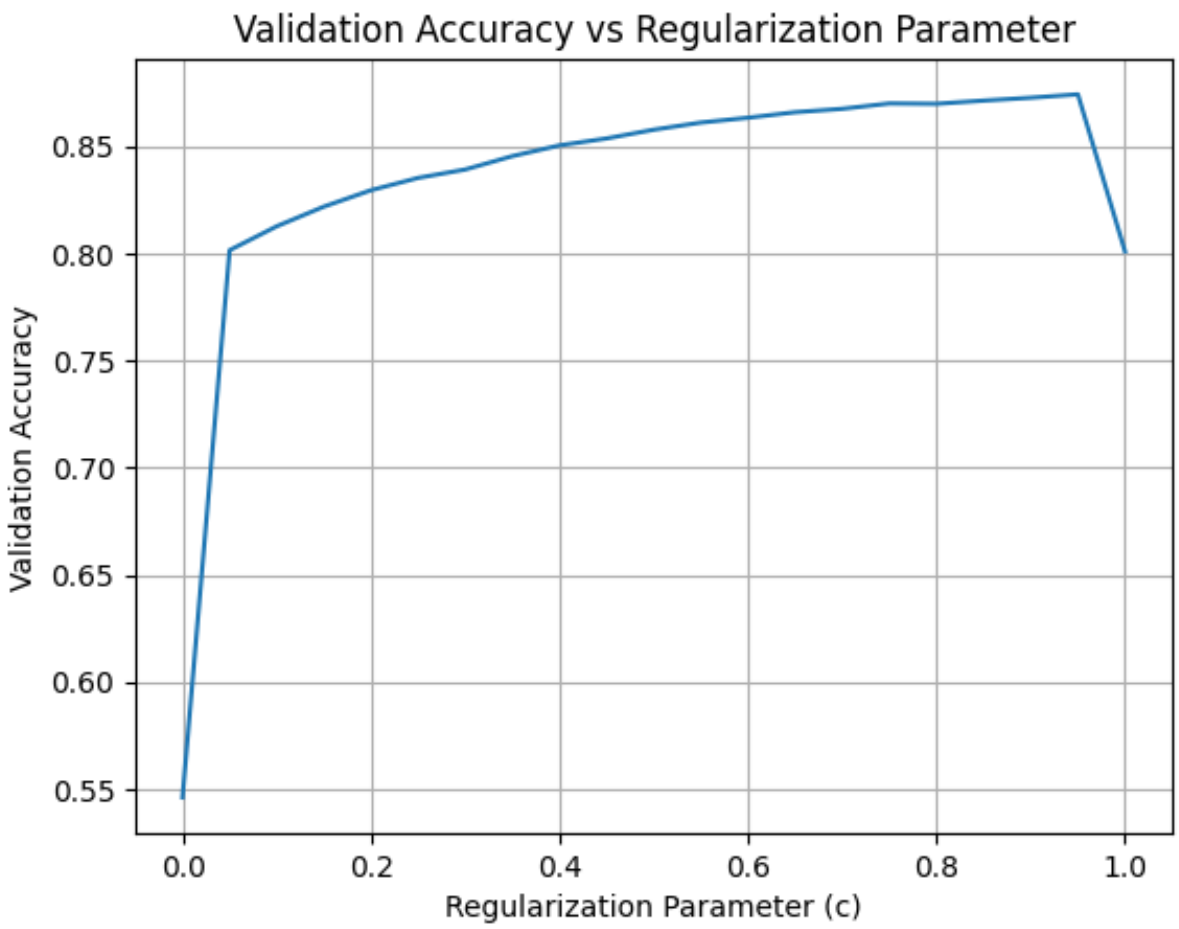


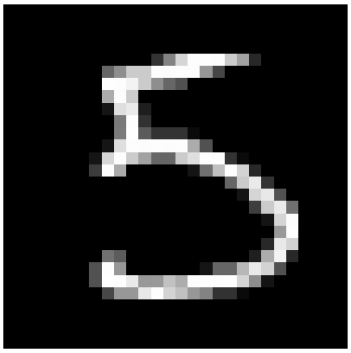

Figure 2: Results from c value testing

**13 (c)**

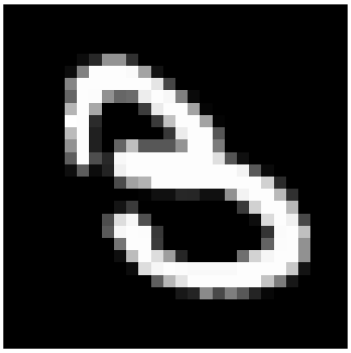
---

- The model with $c = 0.95$ predicted 1178 out of 10,000 incorrectly.
- Therefore, the error rate on the MNIST test set was 0.1178 or 11.78%.
- Note: The accuracy of the model on the MNIST test set (88.22%) was very close to accuracy on the training set(87.42)%.

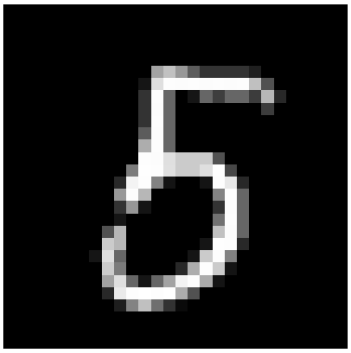
**13 (d)**



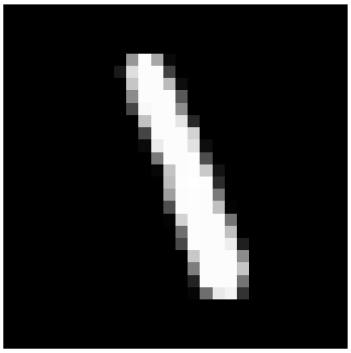
(a) 5 classified as 3



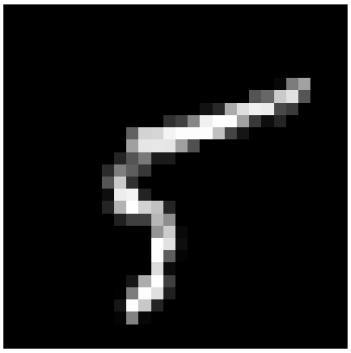
(b) 3 classified as 2



(c) 5 classified as 8



(d) 1 classified as 8



(e) 5 classified as 8

Figure 3: Examples of misclassified digits from the test set