CSCI 5521 (002) Final Exam

Justine John Serdoncillo

TOTAL POINTS

98 / 100

QUESTION 1

Open-ended 25 pts

1.1 (a) **3 / 5**

✓ **- 0 pts** *Correct*

- 5 pts Incorrect solution

- 2.5 pts Incorrect solution for classification

- 2.5 pts Partially incorrect solution

- 2.5 pts Incorrect solution for regression

✓ **- 2 pts** *The explanation for each metric is not enough*

- 2 pts No explanation of the differences between metrics

- 2.5 pts Missing explanation for classification

1.2 (b) **4 / 4**

✓ **- 0 pts** *Correct*

- 2 pts Explanation is not enough

1.3 (c) **4 / 4**

✓ **- 0 pts** *Correct*

- 2 pts Incorrect explanation

1.4 (d) **4 / 4**

✓ **- 0 pts** *Correct*

- 2 pts Explanation is not enough

- 4 pts Explanation is incorrect

1.5 (e) **4 / 4**

✓ **- 0 pts** *Correct*

- 2 pts Explanation is not enough

- 4 pts Answer is missing

1.6 (f) **4 / 4**

✓ **- 0 pts** *Correct*

QUESTION 2

Gaussian Distribution 25 pts

2.1 (a) **10 / 10**

✓ **- 0 pts** *Correct*

- 10 pts Missing

- 3 pts Minor mistake

- 5 pts Major mistake

2.2 (b) **7 / 7**

✓ **- 0 pts** *Correct*

2.3 (c) **8 / 8**

✓ **- 0 pts** *Correct*

- 2 pts Miner mistake

- 4 pts Major mistake

QUESTION 3

Random Forest 25 pts

3.1 (a) **10 / 10**

✓ **- 0 pts** *Correct*

- **2 pts** Miner mistake

3.2 (b) **8 / 8**

✓ **- 0 pts** *Correct*

3.3 (c) **7 / 7**

✓ **- 0 pts** *Correct*

QUESTION 4

SVM 25 pts

4.1 (a) **10 / 10**

✓ **- 0 pts** *Correct*

- **10 pts** Missing

- **5 pts** Major mistake

- **3 pts** Miner mistake

4.2 (b) **7 / 7**

✓ **- 0 pts** *Correct*

- **7 pts** Missing

- **4 pts** Major mistake

- **2 pts** Miner mistake

4.3 (c) **8 / 8**

✓ **- 0 pts** *Correct*

- **8 pts** Missing

Justine John A. Serdoncillo
"JJ"
Final Exam

CSCI 5521: Machine Learning Fundamentals

December 17, 2022

1)

a)

Classification	regression
- Accuracy	- mean absolute error
- top k accuracy	- mean squared error
- f_1 score	- R^2
- log loss	- mean absolute percent error
- precision	
- recall	

- Classification is based on the accuracy of the modeling based on the true and predicted classes which are discreet.
- regression is based on the closeness of the predicted with the actual value. These are not as discreet but are commonly continuos. Both cases can be scaled to the size of the data set.

b) Ensemble methods use sometimes multiple classifiers. They can also introduce randomness which can reduce variance. Additionally by combining diverse different clasifiers it can also reduce bias by being uncorrelated to each other. All of these methods can improve the performance of the model.

c) An estimator is always composed of a variance and bias. Some methods can tweak these two. Depending on the performance of the model the bias and variance can be traded to improve performance. For a model that is underfitting, methods can be done like increasing complexity to reduce bias & increase variance. On the other hand, overfitting can use regularization to increase bias but decrease variance. These methods can improve accuracy.

d) cross-validation can be used by cross-validating the data set to itself. The common one is k-fold where it is split into sets that can be the training and test sets. By doing so, the model can be tested how well it works with unseen data and see its consistency.

e) Machine learning is highly dependent on the model and data. The data quantity and quality should be good enough for the Machine to learn important patterns. A trash model will also produce trash outputs which means that achieving near perfection needs a near perfect model as well. ML is not that easy to understand the inner workings and is hard to generalize to multiple purposes. These limitations can be overcome by proper education and better methods for gathering quality data.

f) Data ethics is really necessary because it keeps ML in check with the purposes and the implications it shows. The model can be used for unethical reasons. At the same time, without further guidance, biases in the world can seep into the model and result to unethical concerns and a closed feedback loop. All these can be ensured by consistent monitoring, diversity and repeated questioning of the effects of the ML algorithms created.

1.1 (a) **3 / 5**

✓ **- 0 pts** *Correct*

- 5 pts Incorrect solution

- 2.5 pts Incorrect solution for classification

- 2.5 pts Partially incorrect solution

- 2.5 pts Incorrect solution for regression

✓ **- 2 pts** *The explanation for each metric is not enough*

- 2 pts No explanation of the differences between metrics

- 2.5 pts Missing explanation for classification

Justine John A. Serdoncillo
"JJ"
Final Exam

CSCI 5521: Machine Learning Fundamentals

December 17, 2022

1)

a)

Classification	regression
- Accuracy	- mean absolute error
- top k accuracy	- mean squared error
- f_1 score	- R^2
- log loss	- mean absolute percent error
- precision	
- recall	

- Classification is based on the accuracy of the modeling based on the true and predicted classes which are discreet.
- regression is based on the closeness of the predicted with the actual value. These are not as discreet but are commonly continous. Both cases can be scaled to the size of the data set.

b) Ensemble methods use sometimes multiple classifiers. They can also introduce randomness which can reduce variance. Additionally by combining diverse different clasifiers it can also reduce bias by being uncorrelated to each other. All of these methods can improve the performance of the model.

c) An estimator is always composed of a variance and bias. Some methods can tweak these two. Depending on the performance of the model the bias and variance can be traded to improve performance. For a model that is underfitting, methods can be done like increasing complexity to reduce bias & increase variance. On the other hand, overfitting can use regularization to increase bias but decrease variance. These methods can improve accuracy.

d) Cross-validation can be used by cross-validating the data set to itself. The common one is k-fold where it is split into sets that can be the training and test sets. By doing so, the model can be tested how well it works with unseen data and see its consistency.

e) Machine learning is highly dependent on the model and data. The data quantity and quality should be good enough for the Machine to learn important patterns. A trash model will also produce trash outputs which means that achieving near perfection needs a near perfect model as well. ML is not that easy to understand the inner workings and is hard to generalize to multiple purposes. These limitations can be overcome by proper education and better methods for gathering quality data.

f) Data ethics is really necessary because it keeps ML in check with the purposes and the implications it shows. The model can be used for unethical reasons. At the same time, without further guidance, biases in the world can seep into the model and result to unethical concerns and a closed feedback loop. All these can be ensured by consistent monitoring, diversity and repeated questioning of the effects of the ML algorithms created.

1.2 (b) **4 / 4**

✓ **- 0 pts** *Correct*

- 2 pts Explanation is not enough

Justine John A. Serdoncillo
"JJ"
Final Exam

CSCI 5521:
Machine Learning
Fundamentals

December 17, 2022

1)

a)

Classification	regression
- Accuracy	- mean absolute error
- top k accuracy	- mean squared error
- f_1 score	- R^2
- log loss	- mean absolute percent error
- precision	
- recall	

- Classification is based on the accuracy of the modeling based on the true and predicted classes which are discreet.
- regression is based on the closeness of the predicted with the actual value. These are not as discreet but are commonly continuos. Both cases can be scaled to the size of the data set.

b) Ensemble methods use sometimes multiple classifiers. They can also introduce randomness which can reduce variance. Additionally by combining diverse different clasifiers it can also reduce bias by being uncorrelated to each other. All of these methods can improve the performance of the model.

c) An estimator is always composed of a variance and bias. Some methods can tweak these two. Depending on the performance of the model the bias and variance can be traded to improve performance. For a model that is underfitting, methods can be done like increasing complexity to reduce bias & increase variance. On the other hand, overfitting can use regularization to increase bias but decrease variance. These methods can improve accuracy.

d) Cross-validation can be used by cross-validating the data set to itself. The common one is k-fold where it is split into sets that can be the training and test sets. By doing so, the model can be tested how well it works with unseen data and see its consistency.

e) Machine learning is highly dependent on the model and data. The data quantity and quality should be good enough for the Machine to learn important patterns. A trash model will also produce trash outputs which means that achieving near perfection needs a near perfect model as well. ML is not that easy to understand the inner workings and is hard to generalize to multiple purposes. These limitations can be overcome by proper education and better methods for gathering quality data.

f) Data ethics is really necessary because it keeps ML in check with the purposes and the implications it shows. The model can be used for unethical reasons. At the same time, without further guidance, biases in the world can seep into the model and result to unethical concerns and a closed feedback loop. All these can be ensured by consistent monitoring, diversity and repeated questioning of the effects of the ML algorithms created.

1.3 (c) **4 / 4**

✓ **- 0 pts** *Correct*

- 2 pts Incorrect explanation

Justine John A. Serdoncillo
"JJ"
Final Exam

CSCI 5521:
Machine Learning
Fundamentals

December 17, 2022

1)

a)

Classification	regression
- Accuracy	- mean absolute error
- top k accuracy	- mean squared error
- f_1 score	- R^2
- log loss	- mean absolute percent error
- precission	
- recall	

- Classification is based on the accuracy of the modeling based on the true and predicted classes which are discreet.
- regression is based on the closeness of the predicted with the actual value. These are not as discreet but are commonly continuos. Both cases can be scaled to the size of the data set.

b) Ensemble methods use sometimes multiple classifiers. They can also introduce randomness which can reduce variance. Additionally by combining diverse different clasifiers it can also reduce bias by being uncorrelated to each other. All of these methods can improve the performance of the model.

c) An estimator is always composed of a variance and bias. Some methods can tweak these two. Depending on the performance of the model the bias and variance can be traded to improve performance. For a model that is underfitting, methods can be done like increasing complexity to reduce bias & increase variance. On the other hand, overfitting can use regularization to increase bias but decrease variance. These methods can improve accuracy.

d) Cross-validation can be used by cross-validating the data set to itself. The common one is k-fold where it is split into sets that can be the training and test sets. By doing so, the model can be tested how well it works with unseen data and see its consistency.

e) Machine learning is highly dependent on the model and data. The data quantity and quality should be good enough for the Machine to learn important patterns. A trash model will also produce trash outputs which means that achieving near perfection needs a near perfect model as well. ML is not that easy to understand the inner workings and is hard to generalize to multiple purposes. These limitations can be overcome by proper education and better methods for gathering quality data.

f) Data ethics is really necessary because it keeps ML in check with the purposes and the implications it shows. The model can be used for unethical reasons. At the same time, without further guidance, biases in the world can seep into the model and result to unethical concerns and a closed feedback loop. All these can be ensured by consistent monitoring, diversity and repeated questioning of the effects of the ML algorithms created.

1.4 (d) **4 / 4**

✓ **- 0 pts** *Correct*

- 2 pts Explanation is not enough

- 4 pts Explanation is incorrect

Justine John A. Serdoncillo
"JJ"
Final Exam

CSCI 5521:
Machine Learning
Fundamentals

December 17, 2022

1)

a)

Classification	regression
- Accuracy	- mean absolute error
- top k accuracy	- mean squared error
- f_1 score	- R^2
- log loss	- mean absolute percent error
- precision	
- recall	

- Classification is based on the accuracy of the modeling based on the true and predicted classes which are discreet.
- regression is based on the closeness of the predicted with the actual value. These are not as discreet but are commonly continuos. Both cases can be scaled to the size of the data set.

b) Ensemble methods use sometimes multiple classifiers. They can also introduce randomness which can reduce variance. Additionally by combining diverse different clasifiers it can also reduce bias by being uncorrelated to each other. All of these methods can improve the performance of the model.

c) An estimator is always composed of a variance and bias. Some methods can tweak these two. Depending on the performance of the model the bias and variance can be traded to improve performance. For a model that is underfitting, methods can be done like increasing complexity to reduce bias & increase variance. On the other hand, overfitting can use regularization to increase bias but decrease variance. These methods can improve accuracy.

d) Cross-validation can be used by cross-validating the data set to itself. The common one is k-fold where it is split into sets that can be the training and test sets. By doing so, the model can be tested how well it works with unseen data and see its consistency.

e) Machine learning is highly dependent on the model and data. The data quantity and quality should be good enough for the Machine to learn important patterns. A trash model will also produce trash outputs which means that achieving near perfection needs a near perfect model as well. ML is not that easy to understand the inner workings and is hard to generalize to multiple purposes. These limitations can be overcome by proper education and better methods for gathering quality data.

f) Data ethics is really necessary because it keeps ML in check with the purposes and the implications it shows. The model can be used for unethical reasons. At the same time, without further guidance, biases in the world can seep into the model and result to unethical concerns and a closed feedback loop. All these can be ensured by consistent monitoring, diversity and repeated questioning of the effects of the ML algorithms created.

1.5 (e) **4 / 4**

✓ **- 0 pts** *Correct*

- 2 pts Explanation is not enough

- 4 pts Answer is missing

Justine John A. Serdoncillo
"JJ"
Final Exam

CSCI 5521:
Machine Learning
Fundamentals

December 17, 2022

1)

a)

Classification	regression
- Accuracy	- mean absolute error
- top k accuracy	- mean squared error
- f_1 score	- R^2
- log loss	- mean absolute percent error
- precision	
- recall	

o Classification is based on the accuracy of the modeling based on the true and predicted classes which are discreet.

o regression is based on the closeness of the predicted with the actual value. These are not as discreet but are commonly continuos. Both cases can be scaled to the size of the data set.

b) Ensemble methods use sometimes multiple classifiers. They can also introduce randomness which can reduce variance. Additionally by combining diverse different clasifiers it can also reduce bias by being uncorrelated to each other. All of these methods can improve the performance of the model.

c) An estimator is always composed of a variance and bias. Some methods can tweak these two. Depending on the performance of the model the bias and variance can be traded to improve performance. For a model that is underfitting, methods can be done like increasing complexity to reduce bias & increase variance. On the other hand, overfitting can use regularization to increase bias but decrease variance. These methods can improve accuracy.

d) Cross-validation can be used by cross-validating the data set to itself. The common one is k-fold where it is split into sets that can be the training and test sets. By doing so, the model can be tested how well it works with unseen data and see its consistency.

e) Machine learning is highly dependent on the model and data. The data quantity and quality should be good enough for the Machine to learn important patterns. A trash model will also produce trash outputs which means that achieving near perfection needs a near perfect model as well. ML is not that easy to understand the inner workings and is hard to generalize to multiple purposes. These limitations can be overcome by proper education and better methods for gathering quality data.

f) Data ethics is really necessary because it keeps ML in check with the purposes and the implications it shows. The model can be used for unethical reasons. At the same time, without further guidance, biases in the world can seep into the model and result to unethical concerns and a closed feedback loop. All these can be ensured by consistent monitoring, diversity and repeated questioning of the effects of the ML algorithms created.

1.6 (f) **4 / 4**

✓ **- 0 pts** *Correct*

2) $\underline{\mathcal{X}} = \{x_1, x_2 \dots x_n\}$ $x_i \in \mathbb{R}$ $\mu \in \mathbb{R}$ $\sigma^2 \in \mathbb{R}_{++}$

a) $$P(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

$$l(\mu, \sigma_n^2 | \mathcal{X}) = \prod_{t=1}^{N} P(\mathcal{X} | \mu, \sigma^2) = \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right]$$

$$\mathcal{L}(\mu, \sigma^2 | \mathcal{X}) = \log l(\mu, \sigma^2 | \mathcal{X}) = -\frac{N}{2}(\log 2\pi) - N \log \sigma - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}$$

$$\frac{\partial \mathcal{L}(\mu, \sigma^2 | \mathcal{X})}{\partial \mu} = -\left(\frac{2\sum_{i=1}^{n} x_i - 2\mu N}{2\sigma^2}\right) = 0$$

$$\cancel{2}\sum_{i=1}^{n} x_i - \cancel{2}\mu N = 0$$

$$\sum_{i=1}^{n} x_i = \mu N$$

$$\mu = \frac{\sum_{i=1}^{n} x_i}{N} \qquad \boxed{\hat{\mu}_n = \frac{\sum_{i=1}^{n} x_i}{N}}$$

$$\frac{\partial \mathcal{L}(\mu, \sigma^2 | \mathcal{X})}{\partial \sigma} = -\frac{N}{\sigma} + \frac{\cancel{2}\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2}{\cancel{2}\sigma^3} = 0$$

$$\frac{N}{\cancel{\sigma}} = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2}{\sigma^{\cancel{3}2}}$$

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2}{N}$$

$$\boxed{\hat{\sigma}_n^2 = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2}{N}}$$

b) $$E[\hat{\mu}_n] = E\left[\frac{\sum_{i=1}^{n} x_i}{N}\right] = \frac{1}{N} E\left[\sum_{i=1}^{n} x_i\right]$$

$$= \frac{1}{N}\sum_{i=1}^{n} E[x_i] = \frac{1}{N} N\mu = \mu$$

$$\boxed{E[\hat{\mu}_n] = \mu}$$ it is true because for an unbiased μ it equals μ

c) $$E[\hat{\sigma}_n^2] = E\left[\frac{\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2}{N}\right]$$

$$= \frac{1}{N}\left(E\left[\sum_{i=1}^{n}(x_i)^2 - N\hat{\mu}_n^2\right]\right)$$

$$= \frac{1}{N}\left(\sum_{i=1}^{n} \underbrace{E[(x_i)^2]}_{A} - N \underbrace{E[\hat{\mu}_n^2]}_{B}\right)$$

by definition

$$Var[x] = E[x^2] - E[x]^2$$

$$\hookrightarrow E[x^2] = Var[x] + E[x]^2$$

(A) $$E[(x_i)^2] = Var[x_i] + E[x_i]^2$$

$$E[(x_i)^2] = \sigma^2 + \mu^2$$

2.1 (a) **10 / 10**

✓ **- 0 pts** *Correct*

- 10 pts Missing

- 3 pts Minor mistake

- 5 pts Major mistake

2) $\underline{\mathcal{X}} = \{x_1, x_2 \dots x_n\}$ $\quad \underline{x_i} \in \mathbb{R} \quad \mu \in \mathbb{R} \quad \sigma^2 \in \mathbb{R}_{++}$

a) $$P(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

$$l(\mu, \sigma_n^2 \mid \mathcal{X}) = \prod_{t=1}^{N} P(\mathcal{X} \mid \mu, \sigma^2) = \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right]$$

$$\mathcal{L}(\mu, \sigma^2 \mid \mathcal{X}) = \log l(\mu, \sigma^2 \mid \mathcal{X}) = -\frac{N}{2}(\log 2\pi) - N\log\sigma - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}$$

$$\frac{\partial \mathcal{L}(\mu, \sigma^2 \mid \mathcal{X})}{\partial \mu} = -\left(\frac{2\sum_{i=1}^{n} x_i - 2\mu N}{2\sigma^2}\right) = 0$$

$$\cancel{2}\sum_{i=1}^{n} x_i - \cancel{2}\mu N = 0$$

$$\sum_{i=1}^{n} x_i = \mu N$$

$$\mu = \frac{\sum_{i=1}^{n} x_i}{N} \qquad \boxed{\hat{\mu}_n = \frac{\sum_{i=1}^{n} x_i}{N}}$$

$$\frac{\partial \mathcal{L}(\mu, \sigma^2 \mid \mathcal{X})}{\partial \sigma} = -\frac{N}{\sigma} + \frac{\cancel{2}\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2}{\cancel{2}\sigma^3} = 0$$

$$\frac{N}{\cancel{\sigma}} = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2}{\sigma^{\cancel{3}2}}$$

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2}{N}$$

$$\boxed{\hat{\sigma}_n^2 = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2}{N}}$$

b) $$E[\hat{\mu}_n] = E\left[\frac{\sum_{i=1}^{n} x_i}{N}\right] = \frac{1}{N} E\left[\sum_{i=1}^{n} x_i\right]$$

$$= \frac{1}{N}\sum_{i=1}^{n} E[x_i] = \frac{1}{N} N\mu = \mu$$

$$\boxed{E[\hat{\mu}_n] = \mu}$$

it is true because for an unbiased μ it equals μ

c) $$E[\hat{\sigma}_n^2] = E\left[\frac{\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2}{N}\right]$$

$$= \frac{1}{N}\left(E\left[\sum_{i=1}^{n}(x_i)^2 - N\hat{\mu}_n^2\right]\right)$$

$$= \frac{1}{N}\left(\underbrace{\sum_{i=1}^{n} E[(x_i)^2]}_{A} - N\underbrace{E[\hat{\mu}_n^2]}_{B}\right)$$

by definition

$$Var[x] = E[x^2] - E[x]^2$$

$$\hookrightarrow E[x^2] = Var[x] + E[x]^2$$

A $$E[(x_i)^2] = Var[x_i] + E[x_i]^2$$

$$E[(x_i)^2] = \sigma^2 + \mu^2$$

2.2 (b) **7 / 7**

✓ **- 0 pts** *Correct*

2) $\underline{\mathcal{X}} = \{x_1, x_2 \dots x_n\}$ $\underline{x_i} \in \mathbb{R}$ $\mu \in \mathbb{R}$ $\sigma^2 \in \mathbb{R}_{++}$

a) $P(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$

$l(\mu, \sigma_n^2 | \mathcal{X}) = \prod_{t=1}^{N} P(\mathcal{X} | \mu, \sigma^2) = \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right]$

$\mathcal{L}(\mu, \sigma_n^2 | \mathcal{X}) = \log l(\mu, \sigma^2 | \mathcal{X}) = -\frac{N}{2}(\log 2\pi) - N \log \sigma - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}$

$\frac{\partial \mathcal{L}(\mu, \sigma^2 | \mathcal{X})}{\partial \mu} = -\left(\frac{2\sum_{i=1}^{n} x_i - 2\mu N}{2\sigma^2}\right) = 0$

$\cancel{2}\sum_{i=1}^{n} x_i - \cancel{2}\mu N = 0$

$\sum_{i=1}^{n} x_i = \mu N$

$\mu = \frac{\sum_{i=1}^{n} x_i}{N}$ $\boxed{\hat{\mu}_n = \frac{\sum_{i=1}^{n} x_i}{N}}$

$\frac{\partial \mathcal{L}(\mu, \sigma^2 | \mathcal{X})}{\partial \sigma} = \frac{-N}{\sigma} + \frac{\cancel{2}\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2}{\cancel{2}\sigma^3} = 0$

$\frac{N}{\cancel{\sigma}} = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2}{\sigma^{\cancel{3}2}}$

$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2}{N}$

$\boxed{\hat{\sigma}_n^2 = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2}{N}}$

b) $E[\hat{\mu}_n] = E\left[\frac{\sum_{i=1}^{n} x_i}{N}\right] = \frac{1}{N} E\left[\sum_{i=1}^{n} x_i\right]$

$= \frac{1}{N}\sum_{i=1}^{n} E[x_i] = \frac{1}{N} N\mu = \mu$

$\boxed{E[\hat{\mu}_n] = \mu}$ it is true because for an unbiased μ it equals μ

c) $E[\hat{\sigma}_n^2] = E\left[\frac{\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2}{N}\right]$

$= \frac{1}{N}\left(E\left[\sum_{i=1}^{n}(x_i)^2 - N\hat{\mu}_n^2\right]\right)$

$= \frac{1}{N}\left(\underbrace{\sum_{i=1}^{n} E[(x_i)^2]}_{A} - N \underbrace{E[\hat{\mu}_n^2]}_{B}\right)$

by definition

$Var[x] = E[x^2] - E[x]^2$

$\hookrightarrow E[x^2] = Var[x] + E[x]^2$

(A) $E[(x_i)^2] = Var[x_i] + E[x_i]^2$

$E[(x_i)^2] = \sigma^2 + \mu^2$

(B) $E[\hat{\mu}_n^2] = Var[\hat{\mu}_n] + E[\hat{\mu}_n]^2$

$Var[\hat{\mu}_n] = Var\left(\frac{\sum_{i=1}^{N} x_i}{N}\right)$

$= \frac{1}{N^2}\sum_{i=1}^{n} Var(x_i) = \frac{1}{N^2}\cancel{N}\sigma^2 = \frac{\sigma^2}{N}$

$E[\hat{\mu}_n^2] = \frac{\sigma^2}{N} + \mu^2$

$E[\hat{\sigma}_n^2] = \frac{1}{\cancel{N}}\left(N(\sigma^2+\mu^2) - \cancel{N}\left(\frac{\sigma^2}{N}+\mu^2\right)\right)$

$= \sigma^2 + \cancel{\mu^2} - \frac{\sigma^2}{N} - \cancel{\mu^2} = \sigma^2\left(1-\frac{1}{N}\right)$

$\boxed{E[\hat{\mu}_n^2] = \sigma^2\left(1-\frac{1}{N}\right) \neq \sigma^2}$

the two are not equal to each other

---

3) a) Random Forests are trained by training multiple decision trees that can be grown by only a few samples of training data or few attributes. The multiple trees are then compared and combined to each other to do a prediction done a test point. The prediction is done using the majority vote of features from the different Decision Trees. The two key hyper-parameters for RFs are

1. Number of decision trees / samples
2. Number of sample m of attributes to train on each node

b)

Pros of Random Forests:

- easy to code w/ fairly high accuracy
- lower chance of overfitting
- Can handle weird data that can mess up training
- still has the advantages of normal decision trees, like getting key features

Cons of Random Forests:

- hard to understand the inner workings or interpret
- can be computationally expensive if generating lots of trees and takes a long time
- extra hyperparameters to tweak

c) If the number of features is a lot more than $m=1$, the decision tree may capture unimportant features which can be a problem because of possible pruning - This makes a lot of useless trees. If there is no pruning, this can still be a problem because in the end it is compiled using a majority vote which won't make sense because $m=1$ which means no other feature to compare with and is purely random.

2.3 (c) **8 / 8**

✓ **- 0 pts** *Correct*

- 2 pts Miner mistake

- 4 pts Major mistake

(B) $E[\hat{\mu}_n^2] = Var[\hat{\mu}_n] + E[\hat{\mu}_n]^2$

$Var[\hat{\mu}_n] = Var\left(\frac{\sum_{i=1}^{N} x_i}{N}\right)$

$= \frac{1}{N^2}\sum_{i=1}^{n} Var(x_i) = \frac{1}{N^{\cancel{2}}}\cancel{N}\sigma^2 = \frac{\sigma^2}{N}$

$E[\hat{\mu}_n^2] = \frac{\sigma^2}{N} + \mu^2$

$E[\hat{\sigma}_n^2] = \frac{1}{\cancel{N}}\left(\cancel{N}(\sigma^2+\mu^2) - \cancel{N}\left(\frac{\sigma^2}{N}+\mu^2\right)\right)$

$= \sigma^2 + \cancel{\mu^2} - \frac{\sigma^2}{N} - \cancel{\mu^2} = \sigma^2\left(1-\frac{1}{N}\right)$

$$\boxed{E[\hat{\mu}_n^2] = \sigma^2\left(1-\frac{1}{N}\right) \neq \sigma^2}$$

the two are not equal to each other

3) a) Random Forests are trained by training multiple decision trees that can be grown by only a few samples of training data or few attributes. The multiple trees are then compared and combined to each other to do a prediction done a test point. The prediction is done using the majority vote of features from the different Decision Trees. The two key hyper-parameters for RFs are

1. Number of decision trees/samples
2. Number of sample m of attributes to train on each node

b)

Pros of Random Forests:

- easy to code w/ fairly high accuracy
- lower chance of overfitting
- can handle weird data that can mess up training
- still has the advantages of normal decision trees, like getting key features

Cons of Random Forests:

- hard to understand the inner workings or interpret
- can be computationally expensive if generating lots of trees and takes a long time
- extra hyperparameters to tweak

c) If the number of features is a lot more than $m=1$, the decision tree may capture unimportant features which can be a problem because of possible pruning - This makes a lot of useless trees. If there is no pruning, this can still be a problem because in the end it is compiled using a majority vote which won't make sense because $m=1$ which means no other feature to compare with and is purely random.

3.1 (a) **10 / 10**

✓ **- 0 pts** *Correct*

- 2 pts Miner mistake

(B) $E[\hat{\mu}_n^2] = Var[\hat{\mu}_n] + E[\hat{\mu}_n]^2$

$$Var[\hat{\mu}_n] = Var\left(\frac{\sum_{i=1}^{N} x_i}{N}\right)$$

$$= \frac{1}{N^2}\sum_{i=1}^{n} Var(x_i) = \frac{1}{N^{\cancel{2}}}\cancel{N}\sigma^2 = \frac{\sigma^2}{N}$$

$$E[\hat{\mu}_n^2] = \frac{\sigma^2}{N} + \mu^2$$

$$E[\hat{\sigma}_n^2] = \frac{1}{\cancel{N}}\left(N(\sigma^2+\mu^2) - \cancel{N}\left(\frac{\sigma^2}{N}+\mu^2\right)\right)$$

$$= \sigma^2 + \cancel{\mu^2} - \frac{\sigma^2}{N} - \cancel{\mu^2} = \sigma^2\left(1-\frac{1}{N}\right)$$

$$\boxed{E[\hat{\mu}_n^2] = \sigma^2\left(1-\frac{1}{N}\right) \neq \sigma^2}$$

the two are not equal to each other

---

3) a) Random Forests are trained by training multiple decision trees that can be grown by only a few samples of training data or few attributes. The multiple trees are then compared and combined to each other to do a prediction done a test point. The prediction is done using the majority vote of features from the different Decision Trees. The two key hyper-parameters for RFs are

1. Number of decision trees / samples
2. Number of sample m of attributes to train on each node

b)

Pros of Random Forests:

- easy to code w/ fairly high accuracy
- lower chance of overfitting
- can handle weird data that can mess up training
- still has the advantages of normal decision trees, like getting key features

Cons of Random Forests:

- hard to understand the inner workings or interpret
- can be computationally expensive if generating lots of trees and takes a long time
- extra hyperparameters to tweak

c) If the number of features is a lot more than $m=1$, the decision tree may capture unimportant features which can be a problem because of possible pruning - This makes a lot of useless trees. If there is no pruning, this can still be a problem because in the end it is compiled using a majority vote which won't make sense because $m=1$ which means no other feature to compare with and is purely random.

3.2 (b) **8 / 8**

✓ **- 0 pts** *Correct*

(B) $E[\hat{\mu}_n^2] = Var[\hat{\mu}_n] + E[\hat{\mu}_n]^2$

$$Var[\hat{\mu}_n] = Var\left(\frac{\sum_{i=1}^{N} x_i}{N}\right)$$

$$= \frac{1}{N^2}\sum_{i=1}^{n} Var(x_i) = \frac{1}{N^{\cancel{2}}}\cancel{N}\sigma^2 = \frac{\sigma^2}{N}$$

$$E[\hat{\mu}_n^2] = \frac{\sigma^2}{N} + \mu^2$$

$$E[\hat{\sigma}_n^2] = \frac{1}{\cancel{N}}\left(\cancel{N}(\sigma^2+\mu^2) - \cancel{N}\left(\frac{\sigma^2}{N}+\mu^2\right)\right)$$

$$= \sigma^2 + \cancel{\mu^2} - \frac{\sigma^2}{N} - \cancel{\mu^2} = \sigma^2\left(1-\frac{1}{N}\right)$$

$$\boxed{E[\hat{\mu}_n^2] = \sigma^2\left(1-\frac{1}{N}\right) \neq \sigma^2}$$

the two are not equal to each other

---

3) a) Random Forests are trained by training multiple decision trees that can be grown by only a few samples of training data or few attributes. The multiple trees are then compared and combined to each other to do a prediction done a test point. The prediction is done using the majority vote of features from the different Decision Trees. The two key hyper-parameters for RFs are

1. Number of decision trees/samples

2. Number of sample m of attributes to train on each node

b)

Pros of Random Forests:

- easy to code w/ fairly high accuracy
- lower chance of overfitting
- Can handle weird data that can mess up training
- still has the advantages of normal decision trees, like getting key features

Cons of Random Forests:

- hard to understand the inner workings or interpret
- can be computationally expensive if generating lots of trees and takes a long time
- extra hyperparameters to tweak

c) If the number of features is a lot more than $m=1$, the decision tree may capture unimportant features which can be a problem because of possible pruning - This makes a lot of useless trees. If there is no pruning, this can still be a problem because in the end it is compiled using a majority vote which won't make sense because $m=1$ which means no other feature to compare with and is purely random.

3.3 (c) **7 / 7**

✓ **- 0 pts** *Correct*

4) $Z = \{(\underline{x}^1, r^1), \dots (\underline{x}^n, r^n)\}$ $\underline{x}^i \in \mathbb{R}^d$ and $r^i \in \{-1, 1\}$

Linear support vector machines (SVMs)

$$\min_{\underline{w}, w_0, \xi^i i=1..n} \frac{1}{2}\|\underline{w}\|^2 + C \sum_{i=1}^{n} \xi^i \text{ such that } r^i(\underline{w}^T \underline{x}^i + w_0) \geq 1 - \xi^i$$

$$\xi^i \geq 0 \quad i = 1 \dots n$$

a) minimize the hinge loss on $(\underline{w}, w_0)$

$$\min_{\underline{w}, w_0} \sum_{i=1}^{n} \max\left(0, 1 - r^i[\underline{w}^T \underline{x}_i + w_0]\right) + \frac{\lambda}{2}\|\underline{w}\|^2$$

let $\xi^i = \max(0, 1 - r^i(\underline{w}^T x_i + w_0))$

$$\Rightarrow \min_{\underline{w}, w_0, \xi^i} \sum_{i=1}^{n} \xi^i + \frac{\lambda}{2}\|w\|^2 = \frac{1}{\lambda}\sum_{i=1}^{n}\xi^i + \frac{1}{2}\|w\|^2$$

let $C = \frac{1}{\lambda}$

$$\Rightarrow \min_{\underline{w}, w_0, \xi_i} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi^i$$

$$\xi^i \geq 1 - r^i[\underline{w}^T \underline{x}_i + w_0]$$

$$\xi^i - 1 \geq -r^i[\underline{w}^T \underline{x}_i + w_0]$$

$$r^i[\underline{w}^T \underline{x}_i + w_0] \geq 1 - \xi^i$$

now this is the same as above but going backwards

w/ $C = \frac{1}{\lambda}$ & $\xi^i = \max(0, 1 - r^i(w^T x_i + w_0))$

to produce a hinge loss optimation problem & a regularization term

b) Pseudo code

① initialize parameters to be used:
- batch size, learning rate, # of epochs or loss tolerance when stopping

② train the SVM per mini batch

I thought mini-batch & Stochastic Gradient Descent is different tho

run all epochs

run all mini batches

done per batch

calculate $y = ?(\underline{w}^T \underline{x}^i + \underline{w}_0)$

if $y \neq r$

keep w

else

calculate gradient

$w = w - (\text{gradient})(\text{learning rate})$

4.1 (a) **10 / 10**

✓ **- 0 pts** *Correct*

- 10 pts Missing

- 5 pts Major mistake

- 3 pts Miner mistake

4) $\underline{Z} = \{(\underline{x}^1, r^1), \dots (\underline{x}^n, r^n)\}$ $\underline{x}^i \in \mathbb{R}^d$ and $r^i \in \{-1, 1\}$

Linear support vector machines (SVMs)

$$\min_{\underline{w}, w_0, \xi^i i=1..n} \frac{1}{2}\|\underline{w}\|^2 + C \sum_{i=1}^{n} \xi^i \text{ such that } r^i(\underline{w}^T \underline{x}^i + w_0) \geq 1 - \xi^i$$

$$\xi^i \geq 0 \quad i = 1 \dots n$$

a) minimize the hinge loss on $(\underline{w}, w_0)$

$$\min_{\underline{w}, w_0} \sum_{i=1}^{n} \max\left(0, 1 - r^i[\underline{w}^T \underline{x}i + w_0]\right) + \frac{\lambda}{2}\|w\|^2$$

let $\xi^i = \max(0, 1 - r^i(\underline{w}^T x i + w_0))$

$$\Rightarrow \min_{\underline{w}, w_0, \xi^i} \sum_{i=1}^{n} \xi^i + \frac{\lambda}{2}\|w\|^2 = \frac{1}{\lambda}\sum_{i=1}^{n}\xi^i + \frac{1}{2}\|w\|^2$$

let $C = \frac{1}{\lambda}$

$$\Rightarrow \min_{\underline{w}, w_0, \xi_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi^i$$

$$\xi^i \geq 1 - r^i[\underline{w}^T \underline{x}i + w_0]$$

$$\xi^i - 1 \geq -r^i[\underline{w}^T \underline{x}i + w_0]$$

$$r^i[\underline{w}^T \underline{x}i + w_0] \geq 1 - \xi^i$$

now this is the same as above but going backwards

w/ $C = \frac{1}{\lambda}$ & $\xi^i = \max(0, 1 - r^i(\underline{w}^T x i + w_0))$

to produce a hinge loss optimization problem & a regularization term

b) Pseudo code

① initialize parameters to be used:
- batch size, learning rate, # of epochs or loss tolerance when stopping

② train the SVM per mini batch

I thought mini-batch & Stochastic Gradient Descent is different tho

run all epochs

run all mini batches

done per batch

calculate $y = ?(\underline{w}^T \underline{x}^i + \underline{w}_0)$

if $y \neq r$

keep w

else

calculate gradient

$w = w - (\text{gradient})(\text{learning rate})$

4.2 (b) **7 / 7**

✓ **- 0 pts** *Correct*

- 7 pts Missing

- 4 pts Major mistake

- 2 pts Miner mistake

c) $d=2 \quad \underline{x}^i = (x_1^i, x_2^i)$

$$r^i = \begin{cases} +1 & \text{if } (x_1^i + 2x_2^i) \le 1 \\ -1 & \text{otherwise} \end{cases}$$

map $\underline{x}^i$ to $\phi(\underline{x}^i)$ and train the linear SVM

$$\phi(\underline{x}^i) = \left[1 \quad x_1^i \quad x_2^i \quad x_1^i x_2^i \quad (x_1^i)^2 \quad (x_2^i)^2\right]^T$$

SVM solution

$$w = \sum_t \alpha^t r^t \phi(x^t)$$

$$g(x) = w^T \phi(x) = \sum_t \alpha^t r^t \phi(x^t)^T \phi(x)$$

$$\phi(x^t)^T \phi(x) = K(x^t, x)$$

$\phi(x)$ can be a highly accurate predictor if the kernel trick works

$$\phi(x^t)^T \phi(x)$$

$$= 1 + x_1^t x_1 + x_2^t x_2 + x_1^t x_2^t x_1 x_2 + x_1^{t^2} x_1^2 + x_2^{t^2} x_2^2$$

this looks close to the vectorial kernel with $q=2$ $\quad K(x^t, x) = (x^{t^T} x + 1)^2$

$$r^i \begin{cases} +1 & 0 \le 1 - (x_1^i + 2x_2^i)^2 \\ -1 & \text{otherwise} \end{cases}$$

$$r^i = \text{sign}\left[1 - (x_1^i + 2x_2^i)^2\right]$$

Linear SVM

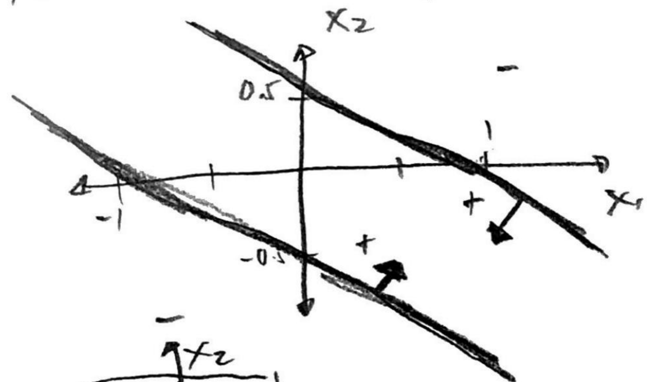
$$\min \frac{1}{2}\|\underline{w}\|^2 + C \sum_t \xi^t$$

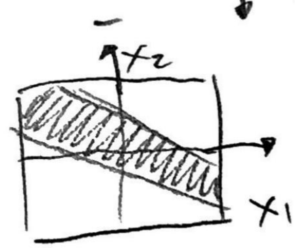
$$r^t(\underline{w}^T \underline{x}^t + \underline{w}_0) \ge 1 - \xi^i$$

$$\sum_t \alpha^t r^t = 0 \quad 0 \le \alpha^t \le \frac{1}{N}$$

$$(x_1 + 2x_2^i)^2 \le 1$$

$$-1 \le x_1 + 2x_2^i \le 1$$ equal to this





given the polynomial kernel of degree 2 and the given boundary of the labels, it is possible that using this mapping can be a highly accurate predictor because it can split it into an outside and inside boundary

Yes

4.3 (C) **8 / 8**

✓ **- 0 pts** *Correct*

- 8 pts Missing