National University of Singapore

CS2109S—Introduction to AI and Machine Learning

Final Exam - Context for Examplify Questions

Semester 1, 2024/2025

Time allowed: 120 minutes

---

1. Please place your student card or identification document (NRIC, driving license, etc.) on the top right-hand corner of your desk.
2. Please switch off your personal devices with communication features and leave them on the floor next to your desk at all times.
3. If you wish to communicate with an invigilator, go to the washroom, or leave before the end of the assessment, please raise your hand to inform the invigilator.
4. Please follow the other instructions in Examplify.

AI (Questions 1-2)

Linear and Logistic Regression (Questions 3-5)

Support Vector Machines (Questions 6-11)

Neural Networks (Questions 12-16)

Context for Questions 15 and 16

Today's neural network architectures frequently contain *skip connections*. These are designs that allow predetermined neural network blocks to be bypassed. These skip connections help with avoiding vanishing gradient problems.

Let us define a toy model for skip connections here.

$$\begin{aligned} \mathbf{f}^{[1]} &= (\mathbf{W}^{[1]})^T \mathbf{x} \\ \mathbf{a}^{[1]} &= g^{[1]}(\mathbf{f}^{[1]}) + \mathbf{x} \\ f^{[2]} &= (\mathbf{W}^{[2]})^T \mathbf{a}^{[1]} \\ \hat{y} &= g^{[2]}(f^{[2]}). \end{aligned}$$

The activation functions and weight matrices are as follows

$$g^{[1]}(x) = g^{[2]}(x) = \sigma(x) = \frac{1}{1+e^{-x}}, \quad \mathbf{W}^{[1]}, \mathbf{W}^{[2]} \in \mathbb{R}^{2\times 2}.$$

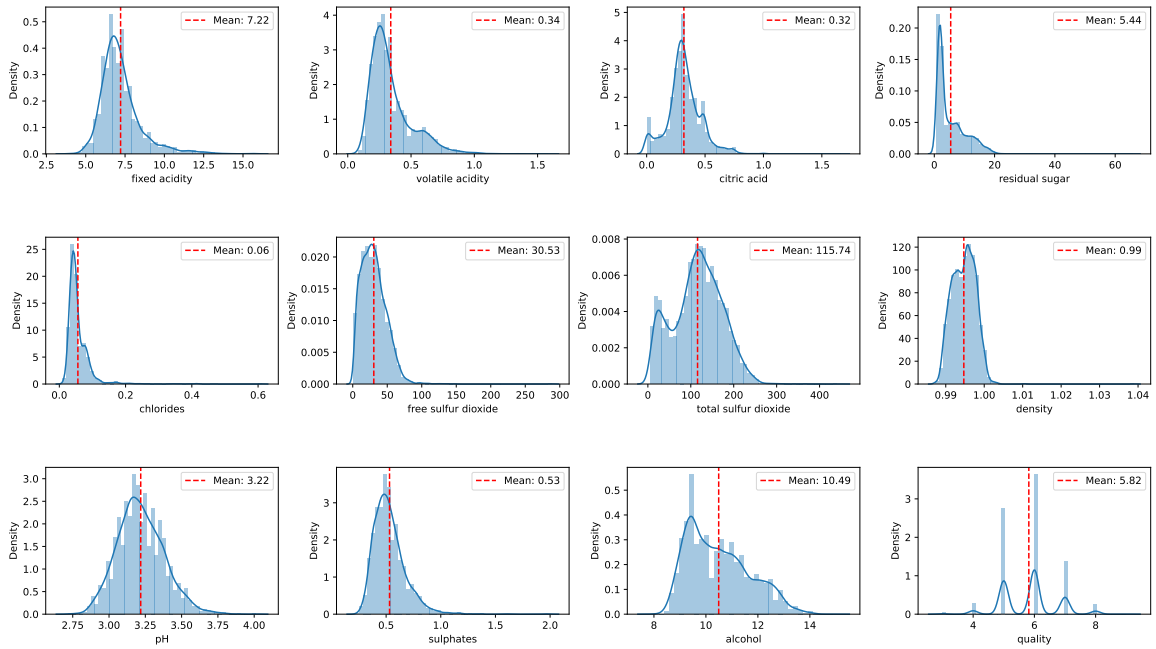
Unsupervised Learning (Questions 17-22)

Case Study: Wine (Questions 23-30)

Wine, a celebrated beverage derived from fermented grapes, is produced when yeast converts grape sugars into ethanol, carbon dioxide, and heat. This dataset, from the UCI Machine Learning Repository, offers an in-depth look into the physicochemical properties that may impact wine quality. It includes comprehensive data on Portuguese "Vinho Verde" wines, with samples from both red (1,599 samples) and white (4,898 samples) varieties combined into a single dataset of 6,497 entries, marked by an additional "type" attribute to differentiate between the two wine types:

- **type**: red or white wine variety
- **fixed acidity**: non-volatile acids contributing to sourness (10 missing values)
- **volatile acidity**: evaporative acids affecting aroma (8 missing values)
- **citric acid**: enhances freshness and balance (3 missing values)
- **residual sugar**: remaining sugar affecting sweetness (2 missing values)
- **chlorides**: salt level influencing taste (2 missing values)
- **free sulfur dioxide**: antimicrobial and antioxidant agent
- **total sulfur dioxide**: preserves wine from spoilage
- **density**: relates to alcohol and sugar content
- **pH**: indicates acidity level (9 missing values)
- **sulphates**: aids preservation and stability (4 missing values)
- **alcohol**: alcohol content by volume
- **quality**: sensory score (0 to 10) reflecting wine quality

A preliminary analysis of the dataset suggests that red and white wine samples share a similar distribution of physicochemical properties and cannot be easily distinguished. Key statistics from the combined dataset are presented below:



Each question for this case study is independent and should be answered based solely on the information provided in the problem description. Use your understanding of the dataset and relevant modeling considerations to select the best answer for each question.

Case Study: DNA (Questions 31-39)

The following case study is inspired from bio-informatics. However, any relationship to actual research projects and actual DNA sequences is co-incidential.

DNA is the carrier of genetic information and modern medicine relies on understanding the role of DNA for combating diseases and designing novel treatments. Think of DNA as follows: DNA is made from the four-letter alphabet G, A, T, C and a DNA sequence is a string made from this alphabet, such as

$$GATACCTTCA \cdots CGGATTA.$$

Duke-NUS has collected anonymized DNA data from 10^4 Singaporeans with their consent. The sequence data obtained are sequences of length 10^6.

Context for Questions 31 and 32

First, your advisor has given you a small subset of the data and provided the following table that holds for consecutive letters in this subset of data. All other options have not been observed in this subset.

Position 1	Position 2	Position 3
T	C	C
A	G	A
G	G	G
T	T	T
C	T	C
G	A	A
A	A	G
C	C	C

Table 1: DNA patterns observed in a small subset of the data

You are building a predictor for the next position based on two previous positions and decide to use a multi-layer Perceptron with a summation in the last layer. You give the two positions the variables x_1 and x_2.