CS344 Exercise 1

Task 0: Survey on the Future of AI (Artificial Intelligence) and ML (Machine Learning)

Complete this ungraded and anonymous survey on the future of AI and ML.

Task 1: Types of Machine Learning (ML) problems

For each example below, circle the type of problem the example corresponds to.

Predicting whether a user will click on an ad or not

Binary classification   Multiclass classification   Regression

Predicting someone's age

Binary classification   Multiclass classification   Regression

Predicting the price of a house

Binary classification   Multiclass classification   Regression

Predicting whether a student applying to college will be accepted, waitlisted, or rejected

Binary classification   Multiclass classification   Regression

Predicting whether or not it's the owner of the phone who's looking at the camera

Binary classification   Multiclass classification   Regression

Task 2: Featurizing (or encoding) data

Many ML systems expect numerical data (i.e., numbers) as input. When we have non-numerical data (e.g., text), we generally convert them to a numerical format.

Binary encoding converts data that pertain to one of two categories (e.g., Yes or No, True or False) to 1 or 0.

Ordinal encoding converts data that pertain to multiple (more than two) categories to integers, one integer per category. Normally, the categories are represented by integers 0, 1, 2, ..., *k*-1, where *k* is the number of different categories. For example, breast cancer ultrasound images that are indicated as normal, benign, or malignant, (*k* = 3) may be represented as 0 (normal), 1 (benign), or 2 (malignant).

One-hot encoding converts data that pertain to multiple (more than two) categories to sequences of binary encodings (0 or 1) where the length of the sequence corresponds to the number of different categories, *k*. For example, personality types described by sanguine, choleric, melancholic, and phlegmatic, (*k*=4) may be represented as 0 0 0 1 (normal), 0 0 1 0 (choleric), 0 1 0 0 (melancholic), and 1 0 0 0 (phlegmatic).

Both ***ordinal encoding*** and ***one-hot encoding*** can be used for data with multiple (more than two) categories and they have different pros and cons. One con of using ***ordinal encoding*** is that it can introduce relationships in the data that don't really exist. For example, suppose we have four flavors of ice cream: mint chocolate chip, rocky road, birthday cake, and Neapolitan. If we use ordinal encoding, 0 (mint chocoloate chip), 1 (rocky road), 2 (birthday cake), and 3 (Neapolitan), we have now unjustifiably caused mint chocolate chip (0) to be closer to rocky road (1) than to Neapolitan (3) because 0 is closer to 1 than it is to 3.

Suppose we have data on movies. Our data contain five features: the year the movie came out (an integer), the movie's content rating (G, PG, PG-13, R, NC-17), whether the movie won an award or not (Award or No Award), the movie's popularity (an integer indicating its Tomatometer score), and the movie's genre (Action, RomCom, Documentary, Horror).
If we converted non-numerical features to numerical features as described above, what would the movie array **(2025, R, Award, 91, RomCom)** be converted to?
Hint: use binary encoding for one feature, ordinal encoding for one feature, and one-hot encoding for one feature. Your resulting array should have length greater than five.

Task 3: Training and testing data

In ML, labeled data are often split into ***training data*** and ***testing data***. ***Training data*** are used to fit a model, learn the model's parameters, and determine all other aspects of data processing and model design. ***Testing data*** are used, once a model and all aspects of the workflow have been finalized, to assess how the model performs on data that the model has never before seen. A common problem in ML is ***data leakage***, which occurs when ***testing data*** are used as ***training data***.

Indicate all options below that are examples of data leakage?

- Should you use ordinal encoding or one-hot encoding on a multi-category feature? To answer this question, you see how well your ML model performs using an ordinal encoding and then using a one-hot encoding and you pick the encoding that yields the best results on the testing data.

- Using some data that your ML model has seen before, you evaluate the model's accuracy. You expect your model will have this level of accuracy going forward on new data that it has never seen before.

- You're not happy with your model's performance on testing data, so you decide to use a completely different type of machine learning model. You assess this new model with the testing data and it performs better than the first model.

- When splitting data, you consider whether 70% or 80% or 90% of the data should be training data (with 30% or 20% or 10% being testing data, respectively). You try all three options and stick with the one that gives the best results on the testing data.

Task 4: Feature scaling

Imagine some college uses two features, high school GPA and standardized test score, to evaluate admission applications (perhaps a poor admission policy!). The college doesn't give the two equal weight, but rather weights GPA much more highly than standard test score. The college calculates a single application score for each applicant by combining GPA and standard test score, where GPA accounts for 90% of the application score and standard test score accounts for 10% of the application score.

Suppose there are two applicants, *A* and *B*, where *A* has a GPA of 3.7 and a standardized test score of 1500, and *B* has a GPA of 1.1 and a standardized test score of 1530.

$A = (3.7, 1500) \qquad B = (1.1, 1530)$

The college must accept exactly one of these two applicants. Which one should be accepted? Given that the college weights GPA more highly than standardized test score, one might expect that the college would choose applicant *A* with the higher GPA. But what are the application scores for each applicant?

Let *w* refer to the weights of each admission feature (90% for GPA and 10% for standard test score):

$w = (0.9, 0.1)$

Applicant *A* has an admission score of: $w \cdot A = 0.9*3.7 + 0.1*1500 \approx 153$
Applicant *B* has an admission score of: $w \cdot B = 0.9*1.1 + 0.1*1530 \approx 154$

B has the higher application score even though *B* has a lower GPA and GPA is given more weight! Why? Because GPAs have smaller magnitude values (between 0.0 and 4.0) than standardized test scores (between 400 and 1600). The two criteria are on different ***scales***.

In ML, we generally don't want the scale of a feature (or the range of values it takes on) to influence its contribution to a final score. So we use ***feature scaling*** to ensure each feature has the same scale. A common ***feature scaling*** approach is to calculate the mean and standard deviation for each feature. Then a scaled version of each feature value is computed by taking the original value, subtracting its feature's mean and then dividing by its feature's standard deviation. This results in each feature having the same scale, i.e., the same mean (0) and the same standard deviation (1).

If $\mu_{\_GPA}$ and $\mu_{\_Test}$ are the means of applicants' GPAs and standardized test scores, respectively, and $\sigma_{\_GPA}$ and $\sigma_{\_Test}$ are the standard deviations of applicants' GPAs and standardized test scores, respectively, then an applicant's scaled GPA and scaled standardized test score would be:

$$GPA_{scaled} = \frac{GPA_{original} - \mu_{\_GPA}}{\sigma_{\_GPA}} \qquad Test_{scaled} = \frac{Test_{original} - \mu_{\_Test}}{\sigma_{\_Test}}$$

For the two applicants *A* and *B*, what is the mean of their GPA’s, μ_{GPA}?
What is the mean of their standardized test scores, μ_{Test}?

For the two applicants *A* and *B*, what is the standard deviaion of their GPA’s, σ_{GPA}?
What is the standard deviation of their standardized test scores, σ_{Test}?

For applicant *A* with original values of (3.7, 1500) for GPA and standardized test score, what are applicant *A*’s feature scaled values?

For applicant *B* with original values of (1.1, 1530) for GPA and standardized test score, what are applicant *B*’s feature scaled values?

Using weights with *w* = (0.9, 0.1), what is *A*’s weighted application score using feature scaled values?
Using weights with *w* = (0.9, 0.1), what is *B*’s weighted application score using feature scaled values?

When ***feature scaling*** data that have been split into ***training*** data and ***testing*** data, do we calculate the mean and standard deviation of each feature from ***all*** data (***training*** and ***testing***)? No! This would result in ***data leakage***. Instead we calculate the mean and standard deviation for each feature based only on the ***training data***. Using these means and standard deviations (calculated only from the ***training data***), we can then feature scale (i.e., transform) the ***training data*** and feature scale (i.e., transform) the ***testing data***.

Suppose applicant *A* and applicant *B* correspond to our ***training data***. We have a third applicant *C*, with a GPA of 3.0 and standardized test score of 1520, that corresponds to our ***testing data***. For applicant *C* with original values of (3.0, 1520) for GPA and standardized test score, what are applicant *C*'s feature scaled values?

Using weights with *w* = (0.9, 0.1), what is *C*'s weighted application score using feature scaled values?

Is applicant *C* more similar (in terms of scaled application score) to applicant *A* or to applicant *B*?

Task 5: Practicing the fundamentals

Download the Jupyter Notebook for Exercise 1 from the course website. Open the Notebook in your web browser and work through it. As you work through the Notebook, answer the following questions.

What is the shape of **`X`**? What is the shape of **`y`**? How many applicants were offered interviews?

What is the shape of **`X_train`**? What is the shape of **`X_test`**? What is the shape of **`y_train`**? What is the shape of **`y_test`**?

What is the model's accuracy on the 2 *testing data* points?

Does the model predict that this new applicant Siiri is offered an interview?

What is the training accuracy on this larger set of data? What is the testing accuracy on this larger set of data?

CS344 Exercise 1 Final Page

In the *TIME* column, please estimate the time you spent on this exercise. Please try to be as accurate as possible; this information will help us to design future exercises.

PART	TIME
Exercise	