ML Concepts - Introduction to ML

Why do we need Machine Learning? What does it do differently?

- ML is the process of training a piece of software, called a model, to make useful predictions or generate content from data.
- Say, we would want to create an app to predict rainfall we could either do it the traditional way or the ML way
- Using a traditional method involves complex physics-based calculations to model weather, while the ML approach we would give an ML model enormous amounts of weather data until the ML model eventually *learned* the mathematical relationship between weather patterns that produce differing amounts of rain. We would then give the model the current weather data, and it would predict the amount of rain.

*What is Supervised Machine Learning? What are the types?

- Supervised learning uses a training set to teach models to yield the desired output. This training dataset includes inputs and correct outputs, which allow the model to learn over time. The algorithm measures its accuracy through the loss function, adjusting until the error has been sufficiently minimized.
- Classification uses an algorithm to accurately assign test data into specific categories. It recognizes specific entities within the dataset and attempts to draw some conclusions on how those entities should be labeled or defined. Common classification algorithms are linear classifiers, support vector machines (SVM), decision trees, k-nearest neighbor, and random forest
- Regression is used to understand the relationship between dependent and independent variables. It is commonly used to make projections, such as for sales revenue for a given business. Linear regression, logistical regression, and polynomial regression are popular regression algorithms.

*Some important terminologies

1. Feature: A **feature** is an input variable
2. Label: A **label** is the thing we're predicting.
3. Model: A **model** defines the relationship between features and label that an ML system uses to make predictions
4. A **labeled example** includes both feature(s) and the label.
5. An **unlabeled example** contains features but not the label
6. **Training** means creating or **learning** the model. That is, you show the model labeled examples and enable the model to gradually learn the relationships between features and label.

7. **Inference** means applying the trained model to unlabeled examples. That is, you use the trained model to make useful predictions (`y'`). For example, during inference, you can predict `medianHouseValue` for new unlabeled examples.

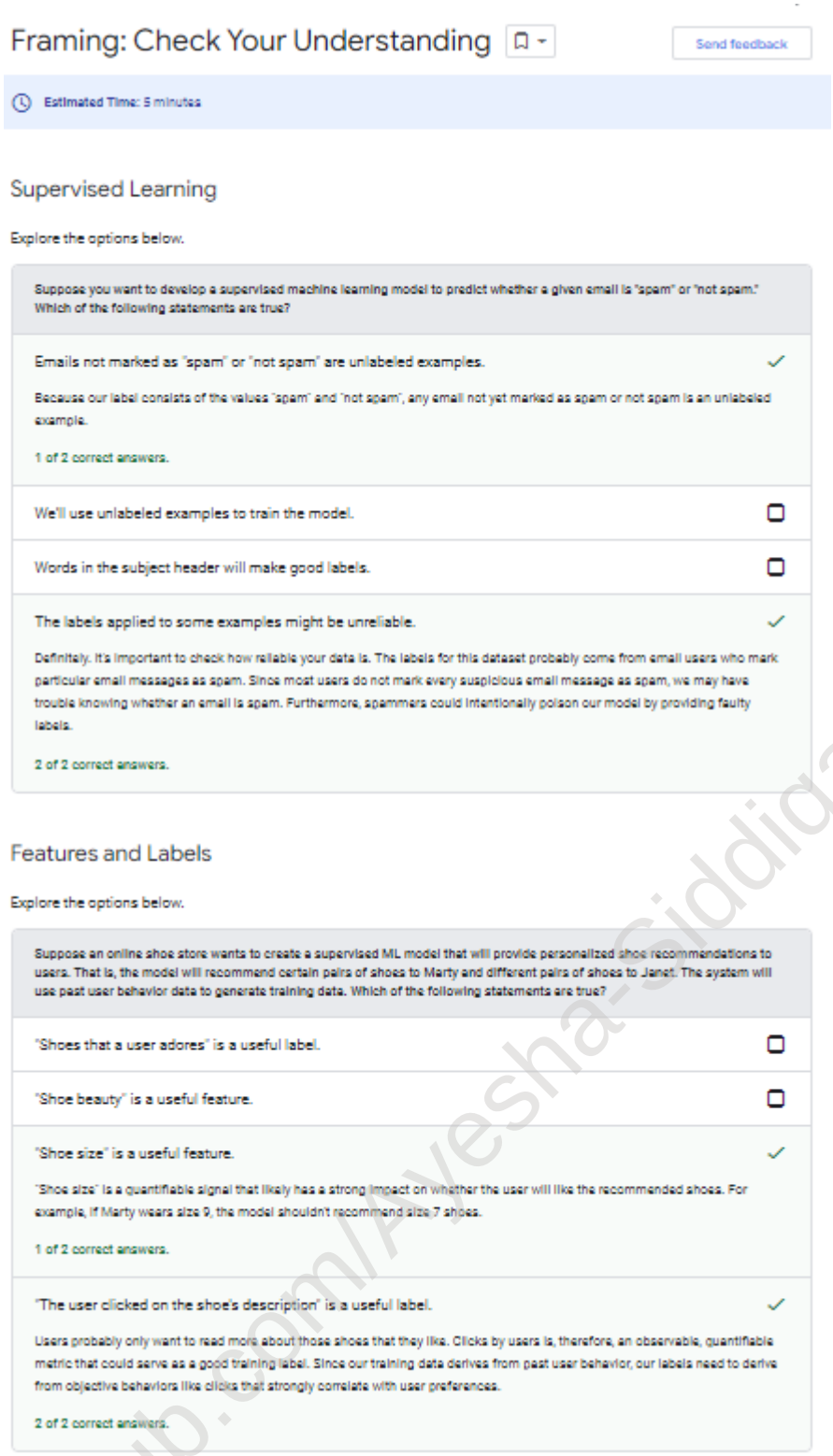
Once we've trained our model with labeled examples, we use that model to predict the label on unlabeled examples

housingMedianAge (feature)	totalRooms (feature)	totalBedrooms (feature)	medianHouseValue (label)
15	5612	1283	66900
19	7650	1901	80100
17	720	174	85700
14	1501	337	73400
20	1454	326	65500

Labeled example

housingMedianAge (feature)	totalRooms (feature)	totalBedrooms (feature)
42	1686	361
34	1226	180
33	1077	271

Unlabeled Example



Framing: Check Your Understanding

Send feedback

Estimated Time: 5 minutes

Supervised Learning

Explore the options below.

Suppose you want to develop a supervised machine learning model to predict whether a given email is "spam" or "not spam." Which of the following statements are true?

- Emails not marked as "spam" or "not spam" are unlabeled examples. ✓
  Because our label consists of the values "spam" and "not spam", any email not yet marked as spam or not spam is an unlabeled example.
  1 of 2 correct answers.
- We'll use unlabeled examples to train the model. ☐
- Words in the subject header will make good labels. ☐
- The labels applied to some examples might be unreliable. ✓
  Definitely. It's important to check how reliable your data is. The labels for this dataset probably come from email users who mark particular email messages as spam. Since most users do not mark every suspicious email message as spam, we may have trouble knowing whether an email is spam. Furthermore, spammers could intentionally poison our model by providing faulty labels.
  2 of 2 correct answers.

Features and Labels

Explore the options below.

Suppose an online shoe store wants to create a supervised ML model that will provide personalized shoe recommendations to users. That is, the model will recommend certain pairs of shoes to Marty and different pairs of shoes to Janet. The system will use past user behavior data to generate training data. Which of the following statements are true?

- "Shoes that a user adores" is a useful label. ☐
- "Shoe beauty" is a useful feature. ☐
- "Shoe size" is a useful feature. ✓
  "Shoe size" is a quantifiable signal that likely has a strong impact on whether the user will like the recommended shoes. For example, if Marty wears size 9, the model shouldn't recommend size 7 shoes.
  1 of 2 correct answers.
- "The user clicked on the shoe's description" is a useful label. ✓
  Users probably only want to read more about those shoes that they like. Clicks by users is, therefore, an observable, quantifiable metric that could serve as a good training label. Since our training data derives from past user behavior, our labels need to derive from objective behaviors like clicks that strongly correlate with user preferences.
  2 of 2 correct answers.

*What is Linear Regression?

A type of machine learning model in which both of the following are true

- The model is a linear model.
- The prediction is a floating-point value. (This is the regression part of linear regression.)
- Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

- This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values.
- There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).
- The relationship can be written as:

$$y = mx + b$$

where:

- y is the temperature in Celsius—the value we're trying to predict.
- m is the slope of the line.
- x is the number of chirps per minute—the value of our input feature.
- b is the y-intercept.

By convention in machine learning, you'll write the equation for a model slightly differently:

$$y' = b + w_1 x_1$$

where:

- y' is the predicted label (a desired output).
- b is the bias (the y-intercept), sometimes referred to as w_0.
- w_1 is the weight of feature 1. Weight is the same concept as the "slope" m in the traditional equation of a line.
- x_1 is a feature (a known input).

To **infer** (predict) the temperature y' for a new chirps-per-minute value x_1, just substitute the x_1 value into this model.

Although this model uses only one feature, a more sophisticated model might rely on multiple features, each having a separate weight (w_1, w_2, etc.). For example, a model that relies on three features might look as follows:

$$y' = b + w_1 x_1 + w_2 x_2 + w_3 x_3$$

*What is a Loss Function?

- Loss function measures the model's prediction accuracy by quantifying the error between predicted and actual outcomes.
- L2 Loss, or squared error, calculates the square of the difference (y - y')² between the predicted value (y') and the true value (y) for a single example i.e. =(observation - prediction)2 = (y - y')2
- The goal of model training is to minimize this loss, aiming to find a set of weights and biases that result in a low average loss across all examples.

$$L_2 Loss = \sum_{(x,y)\in D} (y - prediction(x))^2$$

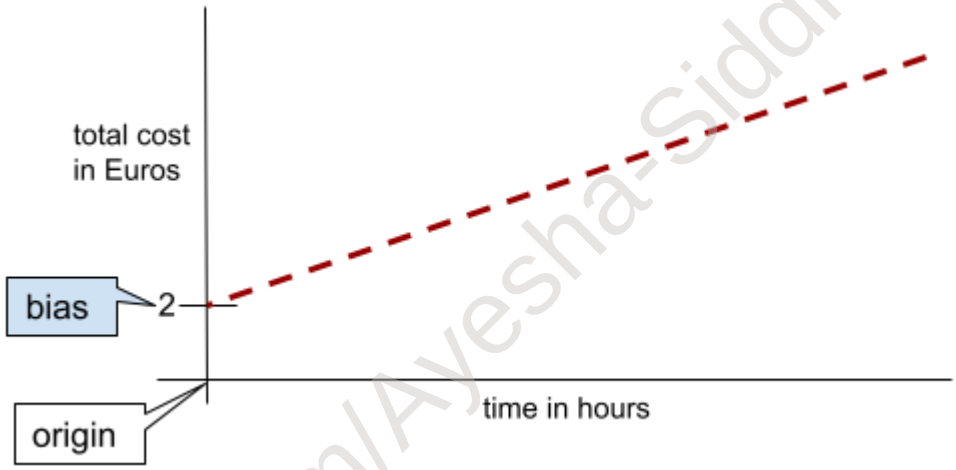
$\sum$:We're summing over all examples in the training set.
D: Sometimes useful to average over all examples, so divide by $\|D\|$.

What is Bias?

- An intercept or offset from an origin. Bias is a parameter in machine learning models, which is symbolized by either of the following:
  - *b*
  - *w0*
- For example, bias is the *b* in the following formula:

$$y' = b + w_1 x_1 + w_2 x_2 + \ldots w_n x_n$$

- In a simple two-dimensional line, bias just means "y-intercept." For example, the bias of the line in the following illustration is 2.



Bias exists because not all models start from the origin (0,0). For example, suppose an amusement park costs 2 Euros to enter and an additional 0.5 Euro for every hour a customer stays. Therefore, a model mapping the total cost has a bias of 2 because the lowest cost is 2 Euros.
Bias is not to be confused with bias in ethics and fairness or prediction bias.

Explain Weights

- A value that a model multiplies by another value. Training is the process of determining a model's ideal weights; inference is the process of using those learned weights to make predictions.
- Imagine a linear model with two features. Suppose that training determines the following weights (and bias):
  - The bias, b, has a value of 2.2
  - The weight, w1 associated with one feature is 1.5.

    - The weight, w2 associated with the other feature is 0.4.
- Now imagine an example with the following feature values:
    - The value of one feature, x1, is 6.
    - The value of the other feature, x2, is 10.
    - This linear model uses the following formula to generate a prediction, y':

$$y' = b + w_1 x_1 + w_2 x_2$$

- Therefore, the prediction is:

$$y' = 2.2 + (1.5)(6) + (0.4)(10) = 15.2$$

- If a weight is 0, then the corresponding feature doesn't contribute to the model. For example, if w1 is 0, then the value of x1 is irrelevant.

What is Mean Squared Loss?

- **Mean square error** (**MSE**) is the average squared loss per example over the whole dataset. To calculate MSE, sum up all the squared losses for individual examples and then divide by the number of examples:

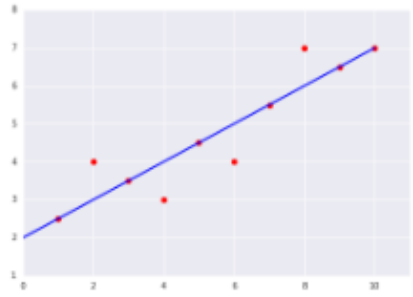
$$MSE = \frac{1}{N} \sum_{(x,y) \in D} (y - prediction(x))^2$$

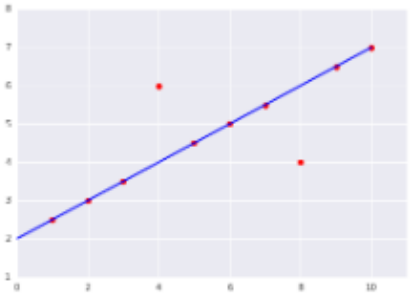
where:

- (x, y) is an example in which
    - x is the set of features (for example, chirps/minute, age, gender) that the model uses to make predictions.
    - y is the example's label (for example, temperature).
- $prediction(x)$ is a function of the weights and bias in combination with the set of features x.
- D is a data set containing many labeled examples, which are (x, y) pairs.
- N is the number of examples in D.

Mean Squared Error

Consider the following two plots:





Explore the options below.

Which of the two data sets shown in the preceding plots has the higher Mean Squared Error (MSE)?

The dataset on the right. ✓

The eight examples on the line incur a total loss of 0. However, although only two points lay off the line, both of those points are *twice* as far off the line as the outlier points in the left figure. Squared loss amplifies those differences, so an offset of two incurs a loss four times as great as an offset of one.

$$MSE = \frac{0^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2}{10} = 0.8$$