

AAIL100: Introduction to Data Science

Ragesh Jaiswal, CSE, IIT Delhi

Inference for Prediction

- Note that our analysis for the linear relationship between variables (response and predictor) has been *descriptive*.
 - Descriptive: *Best-fit line through the scatter plot of the available data items*
- What if the available data items are just a sample from a larger population? Does it make sense to use our sample-based linear predictor for an individual not in our dataset?

Inference for Prediction

- Note that our analysis for the linear relationship between variables (response and predictor) has been *descriptive*.
 - Descriptive: *Best-fit line through the scatter plot of the available data items*
- What if the available data items are just a sample from a larger population? Does it make sense to use our sample-based linear predictor for an individual not in our dataset?
- You realize that there is no way to answer the above question, unless you make certain *assumptions* about the data.
- Assumptions on how data is generated are known as **models**.
- **Regression model**: Assumptions about how data is generated for linearly related variables.

Inference for Prediction

- Note that our analysis for the linear relationship between variables (response and predictor) has been *descriptive*.
 - Descriptive: *Best-fit line through the scatter plot of the available data items*
- What if the available data items are just a sample from a larger population? Does it make sense to use our sample-based linear predictor for an individual not in our dataset?
- You realize that there is no way to answer the above question, unless you make certain *assumptions* about the data.
- Assumptions on how data is generated are known as **models**.
- **Regression model**: Assumptions about how data is generated for linearly related variables.
 - Assumption: The underlying relationship between response and predictor is perfectly linear. The straight line is the **signal** that we must identify from the **noisy** data available to us.
 - What is the assumption about the noise?

Inference for Prediction

- Note that our analysis for the linear relationship between variables (response and predictor) has been *descriptive*.
 - Descriptive: *Best-fit line through the scatter plot of the available data items*
- What if the available data items are just a sample from a larger population? Does it make sense to use our sample-based linear predictor for an individual not in our dataset?
- You realize that there is no way to answer the above question, unless you make certain *assumptions* about the data.
- Assumptions on how data is generated are known as **models**.
- **Regression model**: Assumptions about how data is generated for linearly related variables.
 - Assumption: The underlying relationship between response and predictor is perfectly linear. The straight line is the **signal** that we must identify from the **noisy** data available to us.
 - What is the assumption about the noise?

- **Regression model**:
 - The relation between x and y is perfectly linear. We cannot see this “true line,” but it exists.
 - The scatter plot is created by taking points on the line and pushing them off the line vertically, either above or below, as follows:
 - For each x , find the corresponding point on the true line (that’s the signal), and then generate the noise or error.
 - The errors are drawn at random with replacement from a population of errors that has a normal distribution with mean 0 and constant variance.
 - Create a point whose horizontal coordinate is x and whose vertical coordinate is “the height of the true line at x , plus the error”.
 - Finally, erase the true line from the scatter, and display just the points created.

Inference for Prediction



- **Regression model:**

- The relation between x and y is perfectly linear. We cannot see this “true line,” but it exists.
- The scatter plot is created by taking points on the line and pushing them off the line vertically, either above or below, as follows:
 - For each x , find the corresponding point on the true line (that’s the signal), and then generate the noise or error.
 - The errors are drawn at random with replacement from a population of errors that has a normal distribution with mean 0 and constant variance.
 - Create a point whose horizontal coordinate is x and whose vertical coordinate is “the height of the true line at x , plus the error”.
- Finally, erase the true line from the scatter, and display just the points created.

- **Regression model:**

- The relation between x and y is perfectly linear. We cannot see this “true line,” but it exists.
- The scatter plot is created by taking points on the line and pushing them off the line vertically, either above or below, as follows:
 - For each x , find the corresponding point on the true line (that’s the signal), and then generate the noise or error.
 - The errors are drawn at random with replacement from a population of errors that has a normal distribution with mean 0.
 - Create a point whose horizontal coordinate is x and whose vertical coordinate is “the height of the true line at x , plus the error”.
- Finally, erase the true line from the scatter, and display just the points created.

- How to estimate the slope of the true line from the sample?
- How to estimate the predicted value for an individual not present in the sample?

- **Regression model:**

- The relation between x and y is perfectly linear. We cannot see this “true line,” but it exists.
- The scatter plot is created by taking points on the line and pushing them off the line vertically, either above or below, as follows:
 - For each x , find the corresponding point on the true line (that’s the signal), and then generate the noise or error.
 - The errors are drawn at random with replacement from a population of errors that has a normal distribution with mean 0.
 - Create a point whose horizontal coordinate is x and whose vertical coordinate is “the height of the true line at x , plus the error”.
- Finally, erase the true line from the scatter, and display just the points created.

- How to estimate the slope of the true line from the sample?
- How to estimate the predicted value for an individual not present in the sample?
- **Idea:** Bootstrap
 - Use the bootstrap method to construct confidence intervals on the slope and the predicted values.
 - It works as long as the regression model holds.

Inference for Prediction

- We have used a single feature to predict the outcome.
- Having multiple explanatory features can improve our model's fit to the data and improve predictions.
- We will soon see **multiple linear model** that uses more than one feature to predict the outcome.
 - Instead of $y \approx \theta_0 + \theta_1 x$, we will have $y \approx \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_k x_k$

Classification

- Course content review:
 - Hypothesis testing: *Is the observation due to chance or otherwise?*
 - Parameter Estimation: *Can we estimate some parameter of the population from a sample?*
 - Linear regression: *Can we predict the value of one variable as a linear function of another?*

Classification

- Course content review:
 - Hypothesis testing: *Is the observation due to chance or otherwise?*
 - Parameter Estimation: *Can we estimate some parameter of the population from a sample?*
 - Linear regression: *Can we predict the value of one variable as a linear function of another?*
 - **Classification**: *Given examples with correct **predictions**, can we learn from those examples to make good **predictions** in the future?*
 - Type of prediction: Which **class** does the example belong to?
 - Examples: *Is this online order fraudulent? Does this patient have cancer?* These are examples of **Binary Classification**, where there are only two classes (typically Yes/No).

Classification

- **Classification**: *Given examples with correct **predictions**, can we learn from those examples to make good **predictions** in the future?*
 - Type of prediction: Which **class** does the example belong to?
 - Examples: *Is this online order fraudulent? Does this patient have cancer?* These are examples of **Binary Classification**, where there are only two classes (typically Yes/No).
 - **Definitions**:
 - The individual or situation for which we would like to make a prediction is called an **observation**.
 - Each observation has multiple **attributes**, which are known (e.g., the total value of an online order, a patient's blood report).
 - Each observation has a **class**, which is the answer to the question we care about (e.g., fraudulent or not, cancer or not).
 - We have access to **training data**: a set of observations, each with a known class.
 - The collection of these pre-classified observations is also called a **training set**.
 - A **classification algorithm** is going to analyze the training set, and then come up with a **classifier**: an algorithm for predicting the class of future observations.

Classification

- **Classification**: *Given examples with correct **predictions**, can we learn from those examples to make good **predictions** in the future?*
 - Type of prediction: Which **class** does the example belong to?
 - Examples: *Is this online order fraudulent? Does this patient have cancer?* These are examples of **Binary Classification**, where there are only two classes (typically Yes/No).
 - **Definitions**:
 - The individual or situation for which we would like to make a prediction is called an **observation**.
 - Each observation has multiple **attributes**, which are known (e.g., the total value of an online order, a patient's blood report).
 - Each observation has a **class**, which is the answer to the question we care about (e.g., fraudulent or not, cancer or not).
 - We have access to **training data**: a set of observations, each with a known class.
 - The collection of these pre-classified observations is also called a **training set**.
 - A **classification algorithm** is going to analyze the training set, and then come up with a **classifier**: an algorithm for predicting the class of future observations.

• *Classifiers do not need to be perfect to be useful. They can be useful even if their accuracy is less than 100%.*

Classification

- **Classification**: Given examples with correct *predictions*, can we learn from those examples to make good *predictions* in the future?
- Let's start with ***Binary classification***.
- The simplest binary classifier is based on the *similarity* of a new observation (i.e., with unknown class) with the known observations in the training set.
 - *What is a good measure of similarity?*

Classification

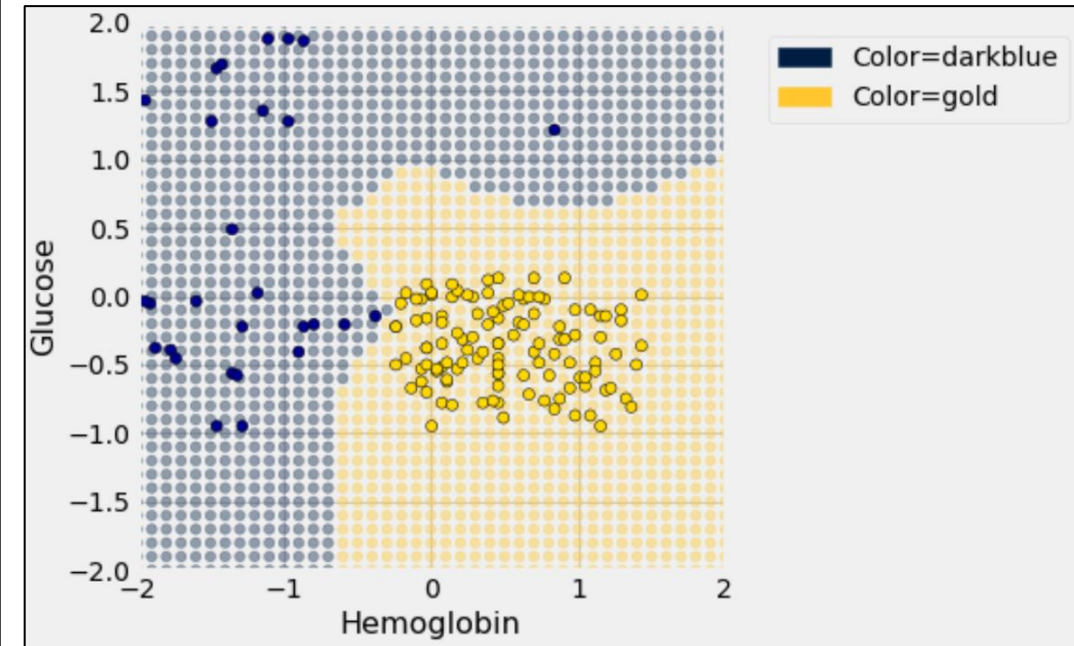
- **Classification**: Given examples with correct *predictions*, can we learn from those examples to make good *predictions* in the future?
- Let's start with ***Binary classification***.
- The simplest binary classifier is based on the *similarity* of a new observation (i.e., with unknown class) with the known observations in the training set.
 - *What is a good measure of similarity?*
 - For numerical attributes: Euclidean distance after conversion to standard units.

Classification

- **Classification**: Given examples with correct *predictions*, can we learn from those examples to make good *predictions* in the future?
- Let's start with ***Binary classification***.
- The simplest binary classifier is based on the *similarity* of a new observation (i.e., with unknown class) with the known observations in the training set.
 - *What is a good measure of similarity?*
 - For numerical attributes: Euclidean distance after conversion to standard units.
 - Classification idea: Find the point p in the training set that is nearest to the new point and predict its class to be the same as that of p .
 - The above idea implicitly defines a ***decision boundary*** for making predictions on new points.

Classification

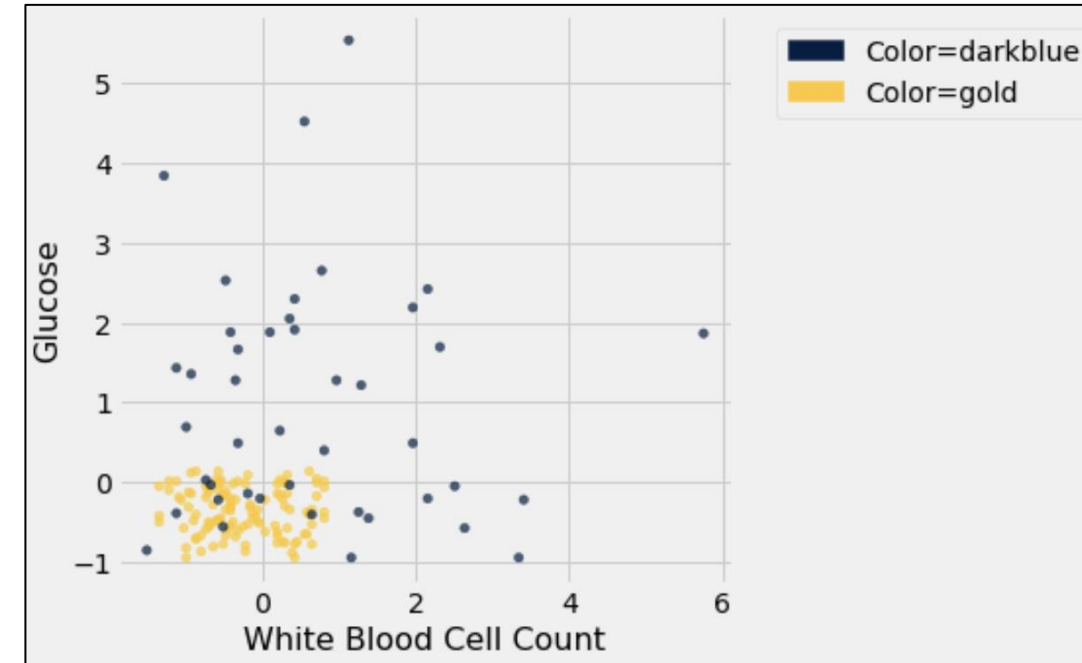
- **Classification**: Given examples with correct *predictions*, can we learn from those examples to make good *predictions* in the future?
- Let's start with **Binary classification**.
- The simplest binary classifier is based on the *similarity* of a new observation (i.e., with unknown class) with the known observations in the training set.
 - *What is a good measure of similarity?*
 - For numerical attributes: Euclidean distance after conversion to standard units.
 - Classification idea: Find the point p in the training set that is nearest to the new point and predict its class to be the same as that of p .
 - The above idea implicitly defines a **decision boundary** for making predictions on new points.



Classification

- **Classification**: Given examples with correct *predictions*, can we learn from those examples to make good *predictions* in the future?
- Let's start with **Binary classification**.
- The simplest binary classifier is based on the *similarity* of a new observation (i.e., with unknown class) with the known observations in the training set.
 - *What is a good measure of similarity?*
 - For numerical attributes: Euclidean distance after conversion to standard units.
 - Classification idea: Find the point p in the training set that is nearest to the new point and predict its class to be the same as that of p .
 - The above idea implicitly defines a **decision boundary** for making predictions on new points.

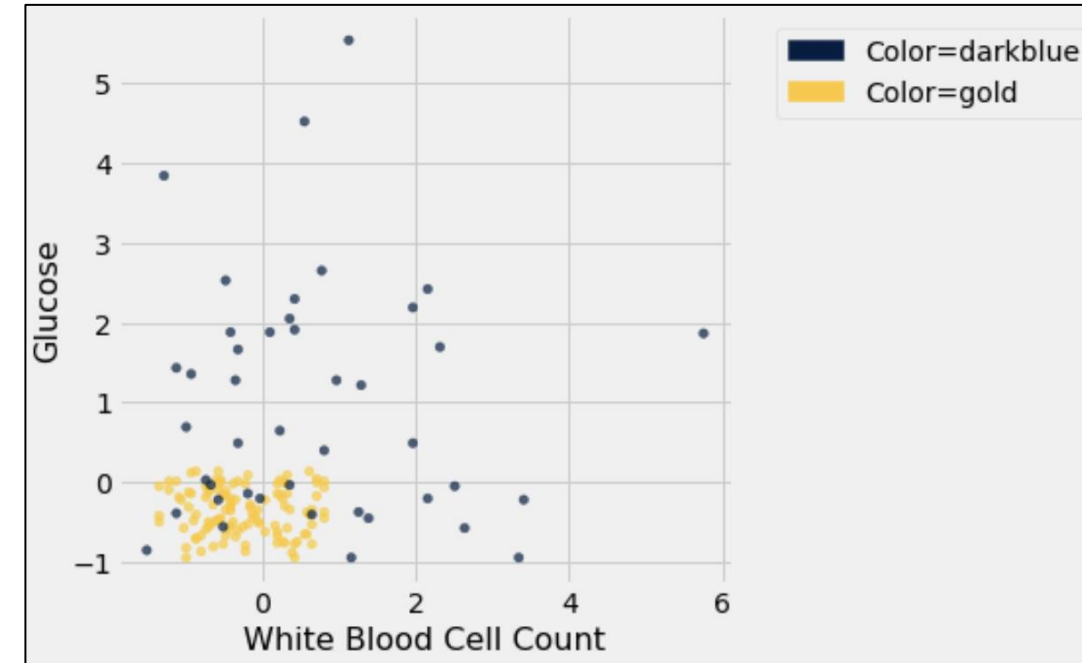
- Sometimes, the decision boundary is not clean



Classification

- **Classification**: Given examples with correct *predictions*, can we learn from those examples to make good *predictions* in the future?
- Let's start with **Binary classification**.
- The simplest binary classifier is based on the *similarity* of a new observation (i.e., with unknown class) with the known observations in the training set.
 - *What is a good measure of similarity?*
 - For numerical attributes: Euclidean distance after conversion to standard units.
 - Classification idea: Find the point p in the training set that is nearest to the new point and predict its class to be the same as that of p .
 - The above idea implicitly defines a **decision boundary** for making predictions on new points.

- Sometimes, the decision boundary is not clean

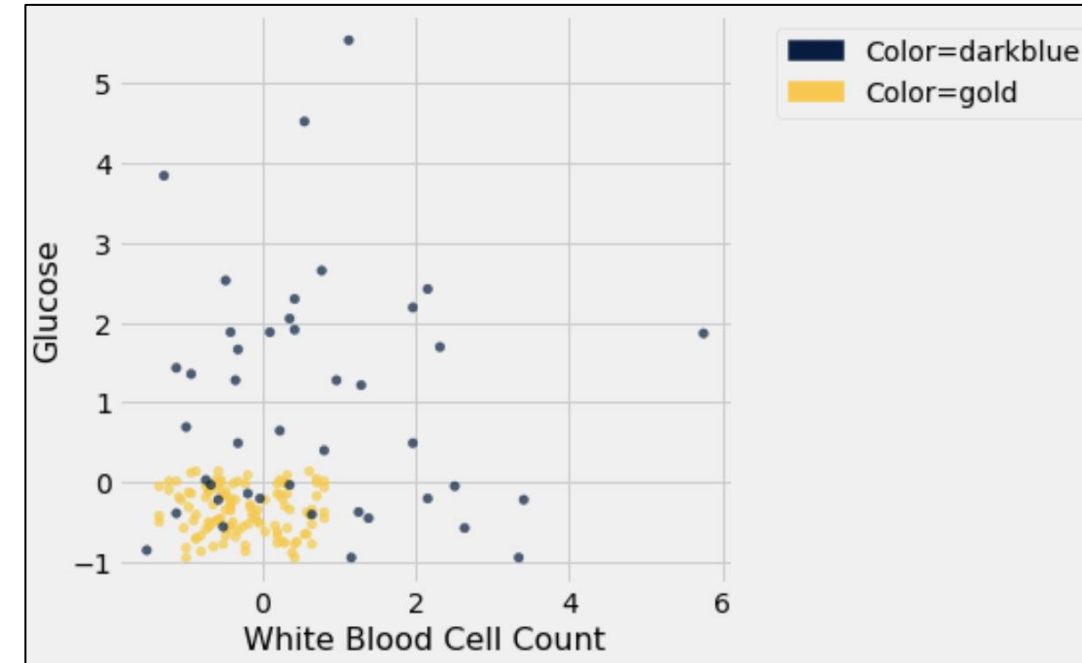


Can you suggest a classification scheme?

Classification

- **Classification**: Given examples with correct *predictions*, can we learn from those examples to make good *predictions* in the future?
- Let's start with **Binary classification**.
- The simplest binary classifier is based on the *similarity* of a new observation (i.e., with unknown class) with the known observations in the training set.
 - *What is a good measure of similarity?*
 - For numerical attributes: Euclidean distance after conversion to standard units.
 - Classification idea: Find the point p in the training set that is nearest to the new point and predict its class to be the same as that of p .
 - The above idea implicitly defines a **decision boundary** for making predictions on new points.

- Sometimes, the decision boundary is not clean



- Idea: Find k nearest neighbors (KNN) and classify based on the majority vote.

End