AWS Course - The Elements of Data Science - part 1

https://www.aws.training/Details/eLearning?id=26598

aws training and certification

Welcome to The Elements of Data Science. In this course, we join Blaine Sundrud for discussions on how to build and continuously improve machine learning models. Topics include the following elements of data science: problem formulation, exploratory data analysis, feature engineering, model training, tuning and debugging, as well as model evaluation and productionizing.

Lesson 2 of 12

What is Data Science?

One goal of Data Science is to uncover actionable insights from seemingly disconnected pieces of information.

What is Data Science?

aws training and certification

- **General Definition:** Processes and systems to extract knowledge or insights from data, either structured or unstructured. *(Wikipedia)*
- **For the purposes of this course:** Managing, analyzing, and visualizing data in support of the Machine Learning workflow.
- But **what is Machine Learning?**

Supervised Learning

Models learn from training data that has been labeled.

Unsupervised Learning

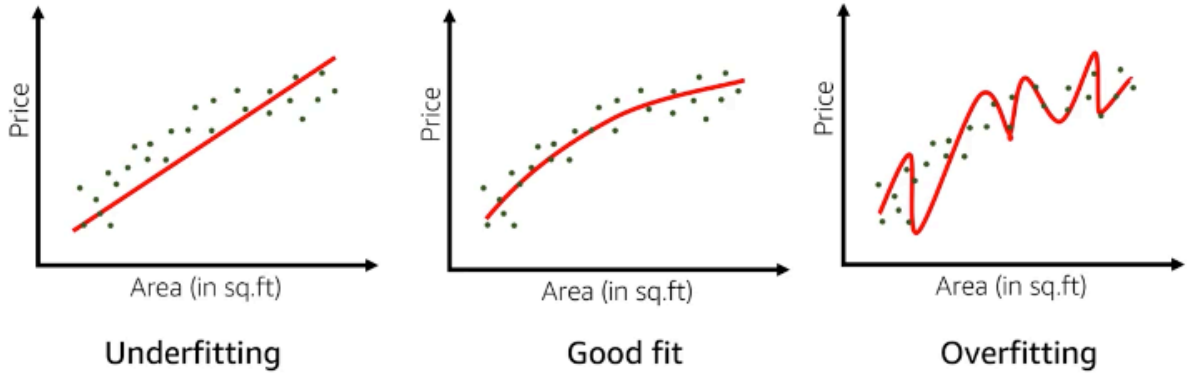
Models learn from test data that has not been labeled.

Reinforcement Learning

Models learn by taking actions that can earn rewards.

Model quality: Underfitting versus overfitting

aws training and certification



Overfitting and underfitting

aws training and certification

Overfitting

- Failure to generalize: Model performs well on training set but poorly on test set
- Typically indicates that model is too flexible for amount of training data
- Flexibility allows it to "memorize" the data, including noise
- Corresponds to high variance – small changes in the training data lead to big changes in the results

Underfitting

- Failure to capture important patterns in the training data set
- Typically indicates model is too simple or there are too few explanatory variables
- Not flexible enough to model real patterns
- Corresponds to high bias – the results show systematic lack of fit in certain regions

Supervised Methods: Linear Regression

In this video we'll discuss linear regression. Topics include linear methods, univariate linear regression, and multivariate linear regression.

Linear methods

aws training and certification

- Parametric methods where function learned has form $f(x) = \phi(w^T x)$ where $\phi()$ is some activation function
- Generally, optimized by learning weights by applying (stochastic) gradient descent to minimize loss function, e.g. $\sum|\hat{y_i} - y_i|^2$
- Simple; a good place to start for a new problem, at least as a baseline
- Methods
  - Linear regression for numeric target outcome
  - Logistic regression for categorical target outcome

Linear regression (univariate)

aws training and certification

- Model relation between a single feature (explanatory variable x) and a real-valued response (target variable y)
- Given data (x, y), and a line defined by w_0 (intercept) and w_1 (slope), the vertical offset for each data point from the line is the error between the true label y and the prediction based on x
- The best line minimizes the **sum of squared errors (SSE)**
- We usually assume the error is Gaussian distributed with mean zero and fixed variance

Linear regression (multivariate)

aws training and certification

- **Multiple linear regression** includes *N* explanatory variables with *N* >= 2:

$$y = w_0x_0 + w_1x_1 + w_2x_2 + \cdots + w_mx_m = \sum_{i=0}^{N} w_ix_i$$

- Sensitive to correlation between features, resulting in high variance of coefficients

- **scikit-learn implementation**: sklearn.linear_model.LinearRegression

Supervised Learning: Logistic Regression and Linear Separability

This video introduces logistic regression and linear separability, and explains how they can be applied to a business problem involving a credit card transaction.

Logistic regression

use sigmoid function: value between 0 and 1 and add probability

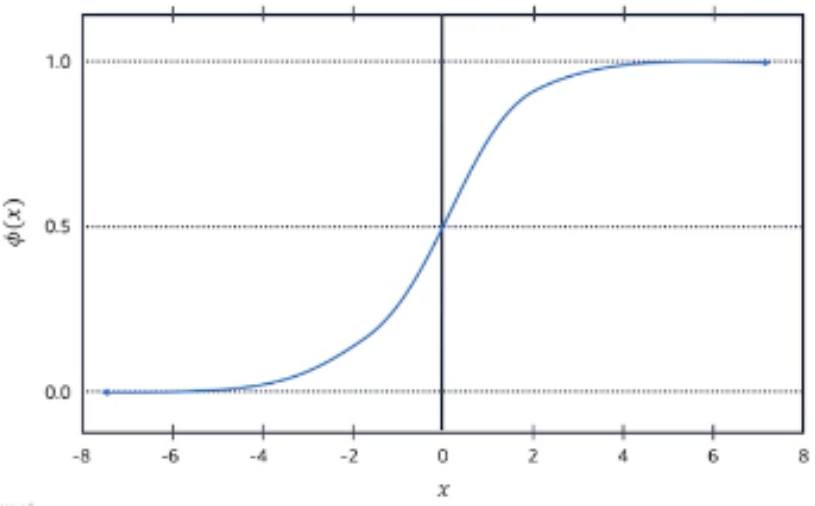
Logistic regression

- Estimates the probability of the input belonging to one of two classes: positive and negative
- Vulnerable to outliers in training data
- scikit-learn: sklearn.linear_model.LogisticRegression
- Relationship to linear model: $\sigma(z) = \frac{1}{1+e^{-z}}$
  - z is a trained multivariate linear function
  - ϕ is a fixed univariate function (not trained)
  - Objective function to maximize = probability of the true training labels

Logistic regression: Sigmoid curve

aws training and certification

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$



Logistic regression

aws training and certification

- Model relation between features (explanatory variables x) and the binary responses (y = 1 or y = 0)
- For all the features, define a linear combination:

$$z = w^T x = w_0 + w_1 x_1 + \cdots + w_n x_n$$

- Define probability of y = 1 given x as p and find the logit of p as

$$logit(p) = log \frac{p}{(1-p)}$$

Logistic regression

- Logistic regression finds the best weight vector by fitting the training data

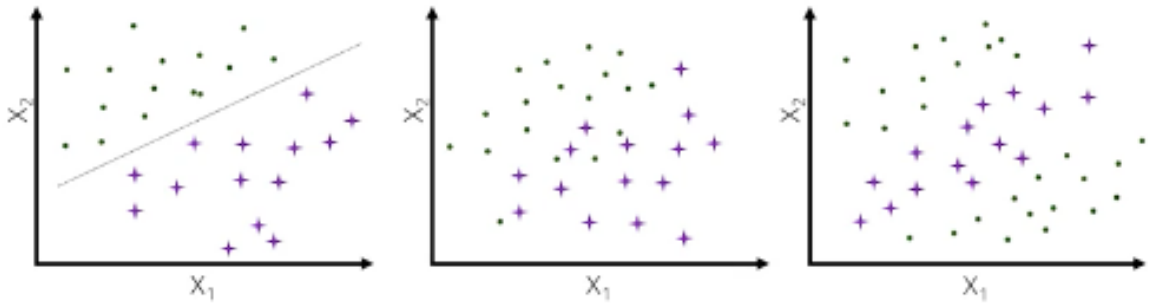
$$logit\big(p(y=1|x)\big) = z$$

- Then, for a new observation, you can use the logistic function $\phi(z)$ to calculate the probability to have label 1. If it is larger than a threshold (for example, 0.5), you will predict the label for the new observation to be positive.

$$z = w^T x = w_0 + w_1 x_1 + \cdots + w_n x_n$$

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

Linearly separable versus non-linearly separable



Linearly separable | Not linearly separable | Not linearly separable

=> cannot apply logistic regression to this pb.

Lesson 3 of 12

Knowledge Check 1

Q1

Which of the following is not a type of machine learning?

- A. Supervised model
- B. Unsupervised model
- C. Semi-supervised model
- D. Reinforcement model
- E. Business Rule model

Q2

You are given a dataset of age, height, weight and gender columns with each row representing a person. Then you are asked to predict the height of a person using the dataset. Drag the following column names to the appropriate category, either labels or features:

Label	Feature
Height	Weight

Q3

I am given a labeled dataset of images where the label is the number zero or one, where zero indicates that the image is not a picture of Jeff Bezos and one indicates that it is. The first thing I try is to train a linear regression model to predict the number zero or one. Is this a good thing to try first?

- A. True
- B. False

Problem Formulation and Exploratory Data Analysis

Problem formulation

aws training and certification

What is the problem you need to solve?

What is the business metric?

Is ML the appropriate approach?

What data is available?

What type of ML problem is it?

What are your goals?

Precisely describe the business problem that you are trying to solve.

Example: Create a way to classify whether the customer credit card transaction is fraudulent.

Problem formulation

aws training and certification

What is the problem you need to solve?

What is the business metric?

Is ML the appropriate approach?

What data is available?

What type of ML problem is it?

What are your goals?

Determine appropriate metrics to measure:

- Quality
- Impact of the solution

Link financial impact with analytics metrics

Example: Rate at which customers are misclassified

Problem formulation

aws training and certification

What is the problem you need to solve?

What is the business metric?

Is ML the appropriate approach?

What data is available?

What type of ML problem is it?

What are your goals?

Can the problem be solved with rules or standard coding?

Are the patterns too difficult to capture algorithmically?

Is there a lot of data available from which to induce patterns?

Problem formulation

aws training and certification

What is the problem you need to solve?

What is the business metric?

Is ML the appropriate approach?

What data is available?

What type of ML problem is it?

What are your goals?

- Summarize the data available.
- Determine the gap between the data you want and the actual data that's available.
- Is there a way to add more data?
- What are the data sources?

Problem formulation

aws training and certification

What is the problem you need to solve?

What is the business metric?

Is ML the appropriate approach?

What data is available?

What type of ML problem is it?

What are your goals?

- Characterize the ML problem according to dimensions.
- Decompose the business problem into a few models.

Problem formulation

aws training and certification

What is the problem you need to solve?

What is the business metric?

Is ML the appropriate approach?

What data is available?

What type of ML problem is it?

What are your goals?

- Establish technical ML goals.
- Define the criteria for successful outcome of the project.

Data Collection

As we've seen, data collection is an ongoing process throughout any machine learning project. This video explores the details of data collection, and open data on AWS.

Data Collection

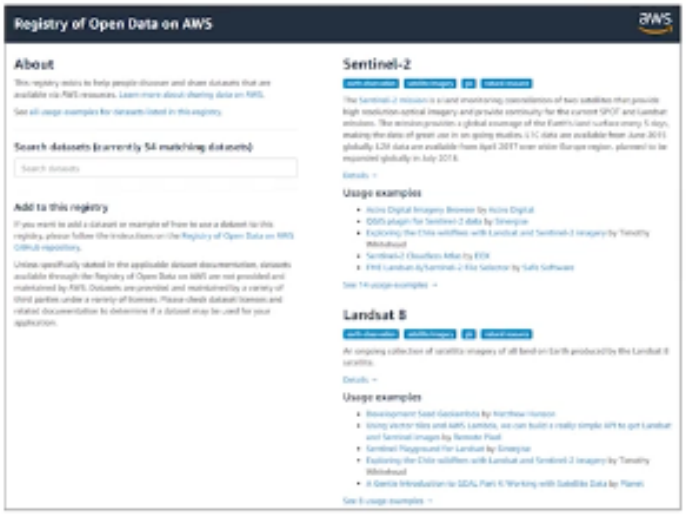
aws training and certification

- Data collection: The process of acquiring training and/or test data
- Motivation:
  - Initial datasets for training models and measuring success
  - Additional data for tuning
  - Replacements for flawed or outdated sets
  - Data and model maintenance post-production
- Sources:
  - Logs
  - Databases
  - Web sites (crawling and scraping)
  - Data providers (public or private)

Registry of Open Data on AWS

aws training and certification

This registry exists to help people **discover** and **share** datasets that are available via AWS resources.



Sampling

Selecting a subset of instances for training and testing

Labeling

Obtaining gold-standard answers for supervised learning

Terminology: Instance = example = data point

Data Collection: Sampling

This video explores two approaches to data collection sampling, and challenges that you should keep in mind when sampling.

Sampling

aws training and certification

Representativity: Sample needs to be representative of the expected production population; i.e., *unbiased*

- Especially important for testing and measurement sets
- It's also important for training sets to get good generalization

Random sampling: Sampling so that each source data point has equal probability of being selected

Stratified Sampling

aws training and certification

Issue: With random sampling, rare subpopulations can be under-represented (or not represented at all).

- A subpopulation is usually defined as examples within the same label.

Stratified sampling: Apply random sampling to each subpopulation separately.

Usually, the sampling probability is the same for each stratum.

- If not, weights can be used in metrics and/or directly in training

Other Issues with Sampling

aws training and certification

Seasonality

Time of day, day of week, time of year (e.g., seasons), holidays, special events, etc.

- Stratified sampling across these can minimize bias
- Visualization can help

Trends

Patterns can shift over time, and new patterns can emerge

- To detect, try comparing models trained over different time periods
- Visualization can help

Consider using validation data that was gathered after your training data was gathered

Leakage

aws training and certification

Train/test bleed: Inadvertent overlap of training and test data when sampling to create datasets

- Sample from fresh data
- Filter out already selected instances
- Partition source data along some dimension (but avoid bias)
- Be especially careful with time-series data and data with duplicate entries

Leakage: Using information during training or validation that is not available in production

- Kaufman, Shachar, *et al.* **Leakage in data mining: Formulation, detection, and avoidance.** *ACM Transactions on Knowledge Discovery from Data* (2012): 15.

Data Collection: Labeling

Here, we review labeling and discuss the use of Amazon Mechanical Turk as a labeling tool.

Labeling

aws training and certification

- Motivation: Often, labels are not readily available in sampled data.
- Examples
  - **Search:** The results a customer wanted to receive
  - **Music categorization:** The genre of a piece
  - **Sentiment analysis:** The overall attitude of the writer
  - **Digitization:** The transcription of handwriting
  - **Object detection:** The localization of objects in images
- Sometimes labels can be inferred (for example, from click-through data)
- Human labels can be preferable to minimize bias, capture subtleties, etc.

Labeling Components

aws training and certification

Labeling guidelines

- Instructions to labelers
- Critical to get right
- Minimize ambiguity

Labeling tools

- Technology:
  - Excel spreadsheets
  - Amazon Mechanical Turk
  - Custom-built tools
- Questions:
  - Human Intelligence Tasks (HITs) should be:
    - Simple
    - Unambiguous

Poor design of either can: 1. Impact labeler productivity and quality, 2. Introduce bias

Labeling Tools

aws training and certification

Amazon Mechanical Turk

- Obtain **human intelligence on demand**
- Access a global, on-demand, 24x7 **workforce**
- Pay only for what you use
- Use for **labeling**

Labeling Tools

aws

Amazon Mechanical Turk

Amazon Mechanical Turk

Sampling and Treatment Assignment

aws training and certification

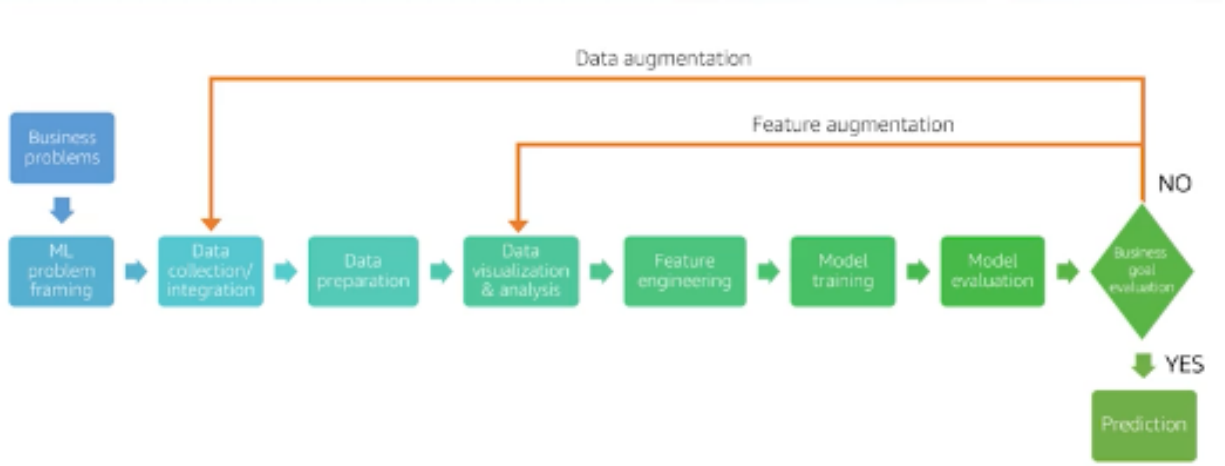
	Random Assignment	No Random Assignment	
Random Sampling	Ideal experiments: Causal conclusion and can be generalized (rarely available in traditional analysis, but become available in online testing)	Typical survey or observation studies: Cannot establish causation, but can find correlation and can be generalized (additional work needed to infer causation)	Generalization
No Random Sampling	Most experiments: Causal conclusion for the sample only (more work is needed to generalize)	Badly-designed survey or pooled studies: Cannot establish causation, cannot generalize to larger population (more work is needed to draw useful conclusions)	No Generalization
	Causation	Correlation	

Exploratory Data Analysis: Domain Knowledge

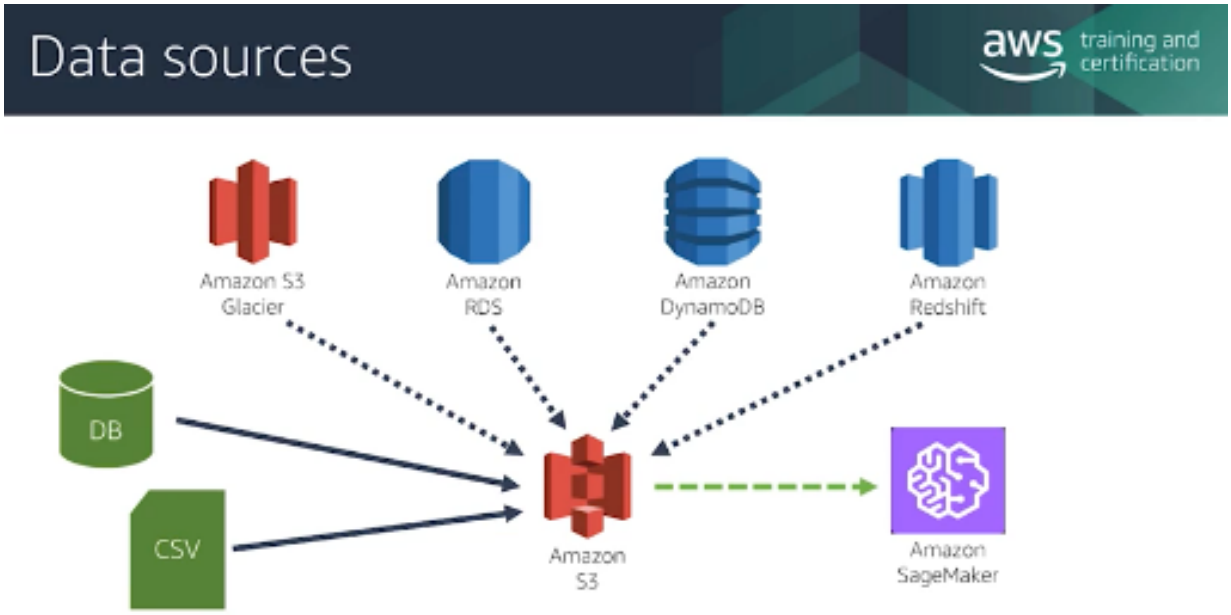
This section explores what needs to happen with data inside a machine learning workflow, in order to produce a successful business solution.

Typical ML Workflow with:
- Data Augmentation

- Feature augmentation



Data Analysis: Data Schema



Ex: breast cancer data

Dataset schema

aws training and certification

*Data Source: Breast Cancer Wisconsin (Diagnostic) Data Set

ID	Diagnosis	Radius mean		Radius std error		Radius worst		Fractal Dim mean
842302	M	17.99	...	1.0950	...	25.380	...	0.11890
842517	M	20.57	...	0.5435	...	24.990	...	0.08902
8510426	B	13.54	...	0.2699	...	15.110	...	0.07259
8510653	B	13.08	...	0.1852	...	14.5	...	0.08183

Label: Radius_mean
Type: Numerical
Range: [6.981-28.11]

Label: Diagnosis
Type: Categorical/Binary
Range: [M=malignant, B=benign]

Label: ID
Type: Text/Numerical
Range: [8670-911320502]

Columns 3-32. Ten real-valued features are computed for each cell nucleus with "mean," "standard error," and "worst:

- Radius (mean of distances from center to points on the perimeter)
- Texture (standard deviation of gray-scale values)
- Perimeter
- Area
- Smoothness (local variation in radius lengths)
- Compactness (perimeter^2/area - 1.0)
- Concavity (severity of concave portions of the contour)
- Concave points (number of concave portions of the contour)
- Symmetry
- Fractal dimension ("coastline approximation" - 1)

Data schema: Merge/joins

aws training and certification

- Types: Categorical, ordinal, numerical, date, vector, text, image, un-structured, etc.
- Use pandas DataFrame merge/join to join two datasets

```
df = pd.DataFrame({"Name":["John Doe","Jim Smith","Elizabeth Shane","David Lee","Hernando Vasquez"],
                   "Job":["Marketing","HR","SDE","SDE","SDE"]})
df_1 = pd.DataFrame({"VP":["Amy Shu","Jane Heart","Ashish Kapoor"],
                   "Job":["SDE","HR","Marketing"]})

df.merge(df_1,on="Job",how = 'inner')
```

	Job	Name	VP
0	Marketing	John Doe	Ashish Kapoor
1	HR	Jim Smith	Jane Heart
2	SDE	Elizabeth Shane	Amy Shu
3	SDE	David Lee	Amy Shu
4	SDE	Hernando Vasquez	Amy Shu

Descriptive Statistics

aws training and certification

Overall statistics

- Number of instances (i.e. number of rows)
- Number of attributes (i.e. number of columns)

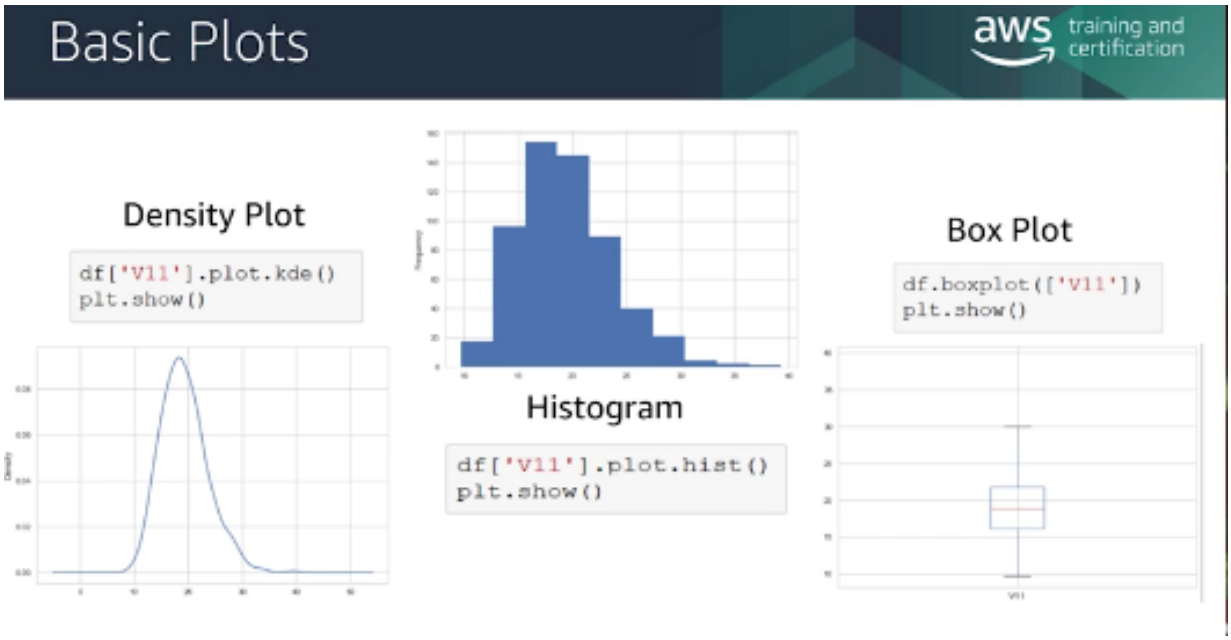
Attribute statistics (univariate)

- Statistics for numeric attributes (mean, variance, etc.) -- df.describe()
- Statistics for categorical attributes (histograms, mode, most/least frequent values, percentage, number of unique values)
  - Histogram of values: E.g., df[<attribute>].value_counts() or seaborn's distplot()
- Target statistics
  - Class distribution: E.g., df[<target>].value_counts() or np.bincount(y)

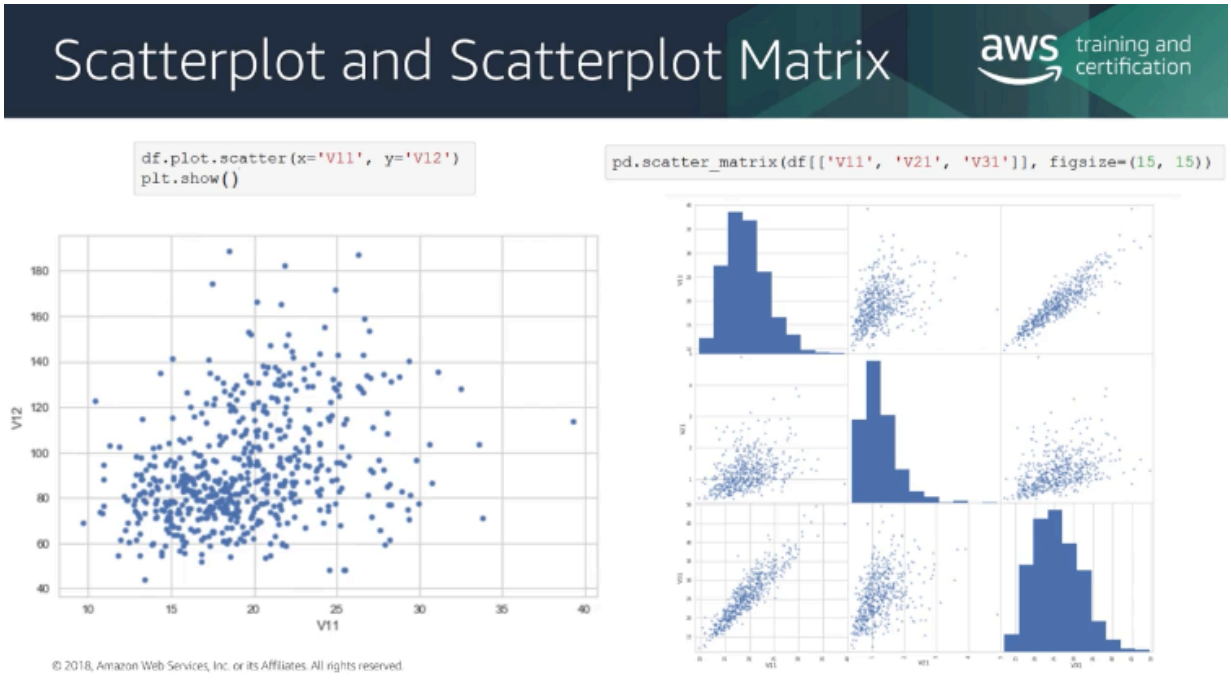
Breast Cancer Classification Problem

aws training and certification

```
import pandas as pd
from sklearn.datasets import load_breast_cancer

dataset = load_breast_cancer()
cols = ['V10', 'V11', 'V12', 'V13', 'V14', 'V15', 'V16', 'V17', 'V18', 'V19',
        'V20', 'V21', 'V22', 'V23', 'V24', 'V25', 'V26', 'V27', 'V28', 'V29',
        'V30', 'V31', 'V32', 'V33', 'V34', 'V35', 'V36', 'V37', 'V38', 'V39']


df = pd.DataFrame(dataset['data'],columns = cols)
df['target'] = dataset.target

# show first a few rows
df.head()
# show datatype for each column
df.info()
# show summary statistics for each column
df.describe()
# check the target variable properties
df['target'].value_counts()
```



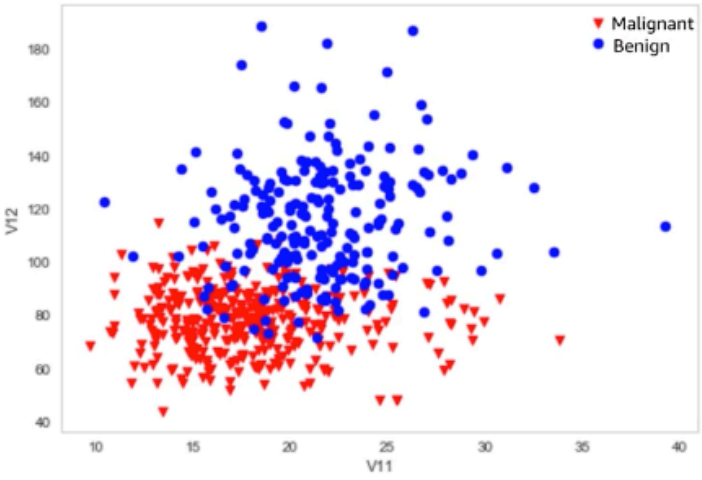
=> plots give an idea of what is inside a feature: peak, range outliers,....



Scatterplot with Identification

aws training and certification

```
malignant = df[['V11', 'V12']][df['target'] == 0]
benign = df[['V11', 'V12']][df['target'] == 1]

plt.scatter(benign['V11'],
            benign['V12'],
            s=50,
            c='red',
            marker='v',
            label='Malignant')

plt.scatter(malignant['V11'],
            malignant['V12'],
            s=50,
            c='blue',
            marker='o',
            label='Benign')
plt.xlabel('V11')
plt.ylabel('V12')
plt.legend()
plt.grid()
plt.show()
```

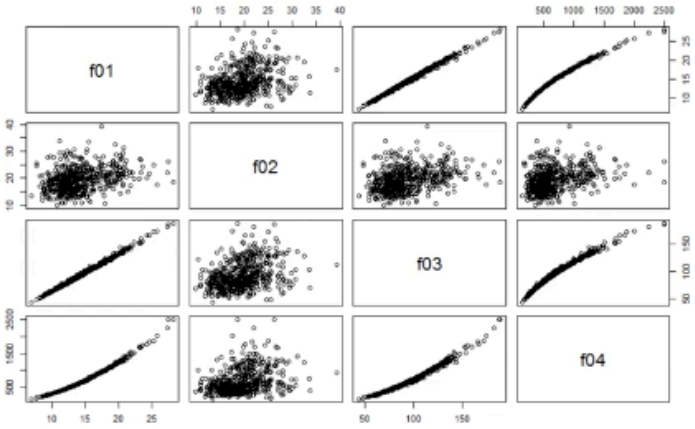


Scatterplot Matrix

aws training and certification

Scatterplot matrices visualize attribute-target and attribute-attribute pairwise relationships

*Data Source: Breast Cancer Wisconsin (Diagnostic) Data Set



Correlation matrix
- highly correlated -> close to 1
- negatively correlated -> negative
- no correlation or little -> close to 0

Correlation Matrix

aws training and certification

*Data Source: Breast Cancer Wisconsin (Diagnostic) Data Set

Correlation matrices measure the linear dependence between features; can be visualized with heat-maps

Example: First 10 features in breast cancer dataset

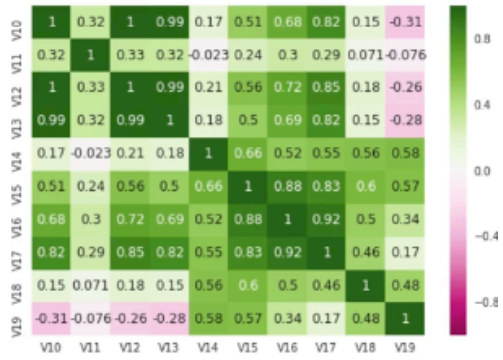
	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19
V10	1.00	0.32	1.00	0.99	0.17	0.51	0.68	0.82	0.15	-0.31
V11	0.32	1.00	0.33	0.32	-0.02	0.24	0.30	0.29	0.07	-0.08
V12	1.00	0.33	1.00	0.99	0.21	0.56	0.72	0.85	0.18	-0.26
V13	0.99	0.32	0.99	1.00	0.18	0.50	0.69	0.82	0.15	-0.28
V14	0.17	-0.02	0.21	0.18	1.00	0.66	0.52	0.55	0.56	0.58
V15	0.51	0.24	0.56	0.50	0.66	1.00	0.88	0.83	0.60	0.57
V16	0.68	0.30	0.72	0.69	0.52	0.88	1.00	0.92	0.50	0.34
V17	0.82	0.29	0.85	0.82	0.55	0.83	0.92	1.00	0.46	0.17
V18	0.15	0.07	0.18	0.15	0.56	0.60	0.50	0.46	1.00	0.48
V19	-0.31	-0.08	-0.26	-0.28	0.58	0.57	0.34	0.17	0.48	1.00

A Matrix heatmap helps us better visualize the correlation

Correlation Matrix Heatmap

aws training and certification

```
col = ['V10', 'V11', 'V12', 'V13', 'V14', 'V15', 'V16', 'V17', 'V18', 'V19']
heatmap = np.corrcoef(df[col].values.T)

fig, ax = plt.subplots(figsize=(15,15))
im = ax.imshow(heatmap, cmap='PiYG',vmin=-1)
fig.colorbar(im)
ax.grid(False)
[[ax.text(j, i, round(heatmap[i, j],2), ha="center", va="center", color="w")
  for j in range(len(heatmap))] for i in range(len(heatmap))]

ax.set_xticks(np.arange(len(col)))
ax.set_yticks(np.arange(len(col)))
ax.set_xticklabels(col)
ax.set_yticklabels(col)

plt.show()
```



*Data Source: Breast Cancer Wisconsin (Diagnostic) Data Set

Seaborn also helps get that easily

Correlation Matrix Heatmap Using Seaborn

```
import seaborn as sns
sns.heatmap(heatmap, yticklabels=col, xticklabels=col,cmap='PiYG',annot=True)
```

*Data Source: Breast Cancer Wisconsin (Diagnostic) Data S

Square root of variance = standard deviation

Formulas for Correlations

Pearson Correlation $$\rho_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_x)^2 \sum_{i=1}^{N}(y_i - \mu_y)^2}}$$

Variance: $\sigma_x^2 = \frac{1}{n}\sum_{i=n}^{n}(x_i - \mu_x)^2$

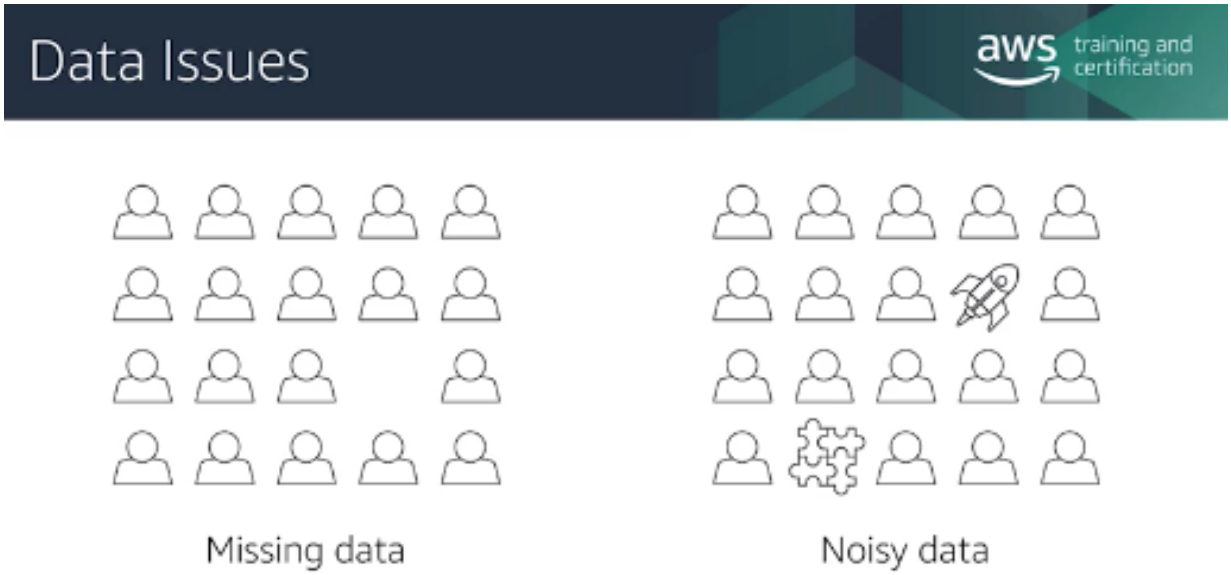
Covariance between (x, y): $\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$

Correlation between (x, y): $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

Exploratory Data Analysis: Data Issues

As previously mentioned, it's important to understand the data you're feeding your machine learning model. In this video we'll discuss how to think about messy and/or missing data.

- Data Issues
  - Messy data
  - Noisy data
  - Biased data
  - Imbalanced data
  - Correlated data



Data Issues

aws training and certification

ID	Survey Response
1	This is grrreat!
2	This is grrreat!
3	Y E s
4	ur $0 L33t!
5	¿Qué?
6	或者
7	احب ذلك

Messy data

ID	Length
1	40 km
2	100 m
3	100 mi
4	74 m
5	74 ft
6	29 in
7	1092 nm

Data on different scales

ID	Measurement
1	96 km
2	twenty
3	5:40:27
4	735 cm^3
5	2 cats
6	44 °C
7	346 Mb/s

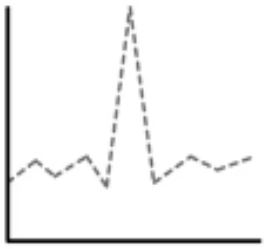
Mixed type data

Data Issues

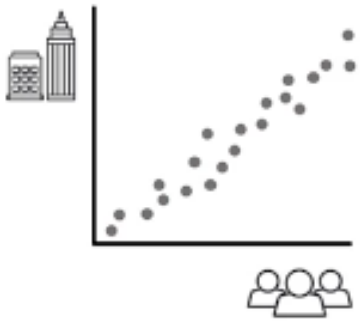
aws training and certification

Imbalanced data

Sample bias



Outliers



Highly correlated features can cause collinearity problems and numerical instability

Lesson 5 of 12

Knowledge Check 2

Q1

If you are looking at a correlation matrix like a heatmap for a 5 numerical-value column dataset. What is the size of the generated matrix?

- A. 5x5
- B. 1x5
- C. 5x1
- D. 1x1

Q2

Which of the following is not a data cleaning issue?

- A. Missing value
- B. Outlier
- C. Correct spelling errors for text data
- D. Quadratic transform of numerical data

Q3

What's the main motivation for using ***stratified*** cross-validation as opposed to simple random cross-validation?

- A. Increase sample size
- B. Decrease sample size
- C. Ensure all sub-groups are represented
- D. Increase computation efficiency