5 - Prática: Estatística p/ Aprendizado de Máquina (I)

Descrição da Atividade

1.Types of Data (Numerical, Categorical, Ordinal)

- **Numerical**
  - **Quantitative Measurement:** Height of people, stock prices...
  - **Discrete Data:** count of some event, how many purchases did a customer make in a year?
  - **Continuous Data**: Infinite number of possible values, how much rain fell on a given day?
- **Categorical**
  - **Qualitative data no inherent mathematical meaning:** gender, binary data, political party.
  - You can assign numbers to categories, but the numbers don't have mathematical meaning.
- **Ordinal**
  - **Mixture of Numerical and Categorical:** has mathematical meaning, movie ratings, 1-5 scale.

2.Mean, Median, Mode

- **Mean**
  - AKA Average.
  - Sum/ number of samples.
- **Median**

    - Sort the Values, and take the value at the midpoint.
    - Is less susceptible to outliers.
- **Mode**
    - The most common value in a data set.

3.Using mean, median, and mode in Python

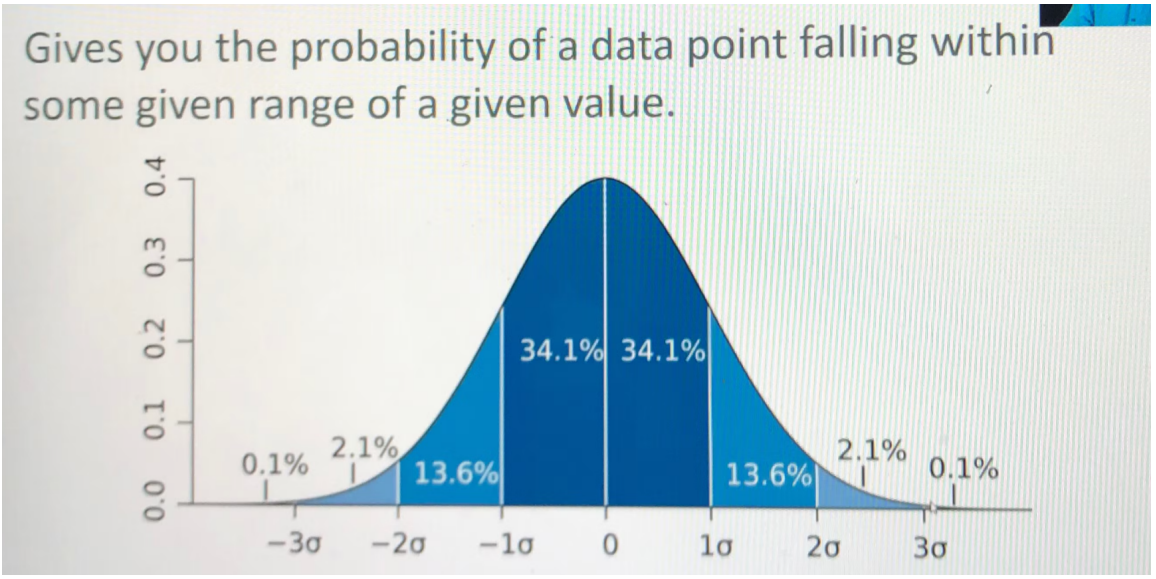
- Notebook Practice

4.Variation and Standard Deviation

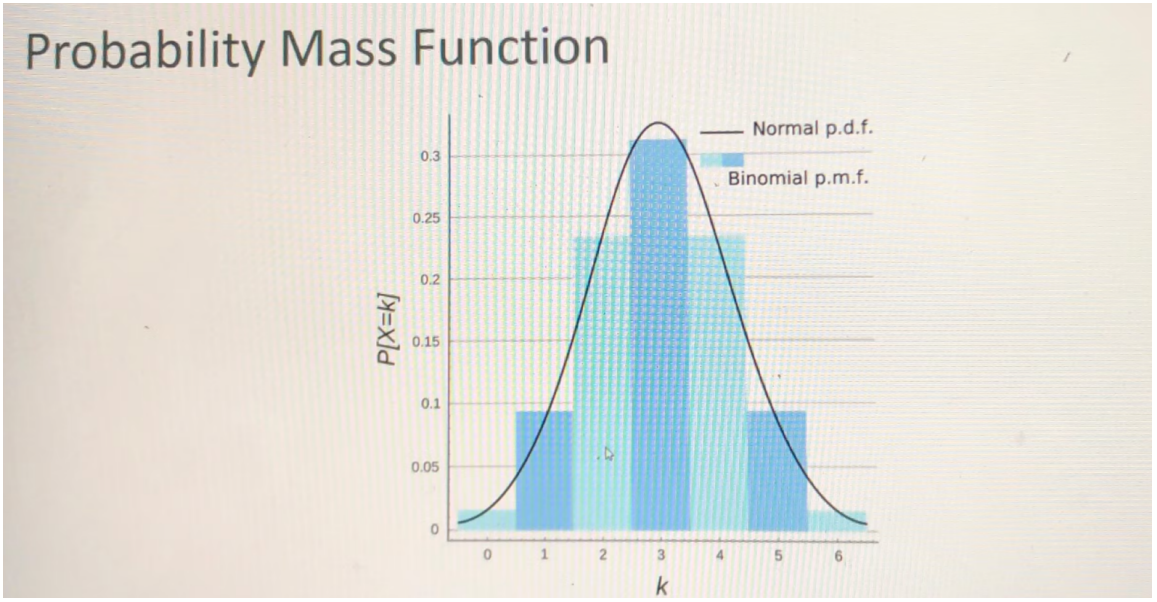
- **Population Variance**
    - Average of the squared differences from the mean.
    - Example: DataSet: 1,4,5,4,8 - Mean: 4.4 - Difference:-3.4, -0.4, 0.6, -0.4, 3.6 - Squared Differences: 11.56, 0.16, 0.36, 0.16, 12.96. Average of the squared diff: 5.04.
- **Sample Variance**
    - 11.56, 0.16, 0.36, 0.16, 12.96/ (4 or N-1) = 6.3.
- **Standard Deviation**
    - Average of the variance: 5.04 square root = 2.24.
    - A way to identify outliers.

5.Probability Density Function; Probability Mass Function

- **Probability Density Function**



- **Probability Mass Function**

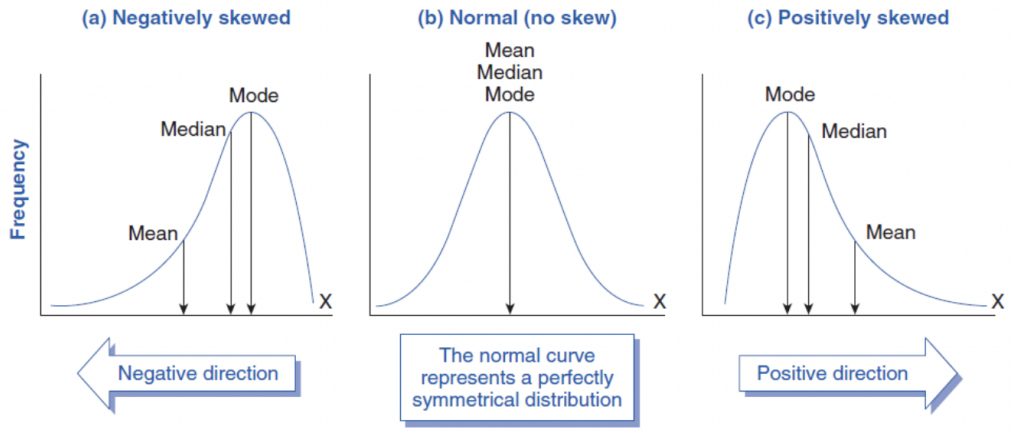


6.Common Data Distributions (Normal, Binomial, Poisson, etc)

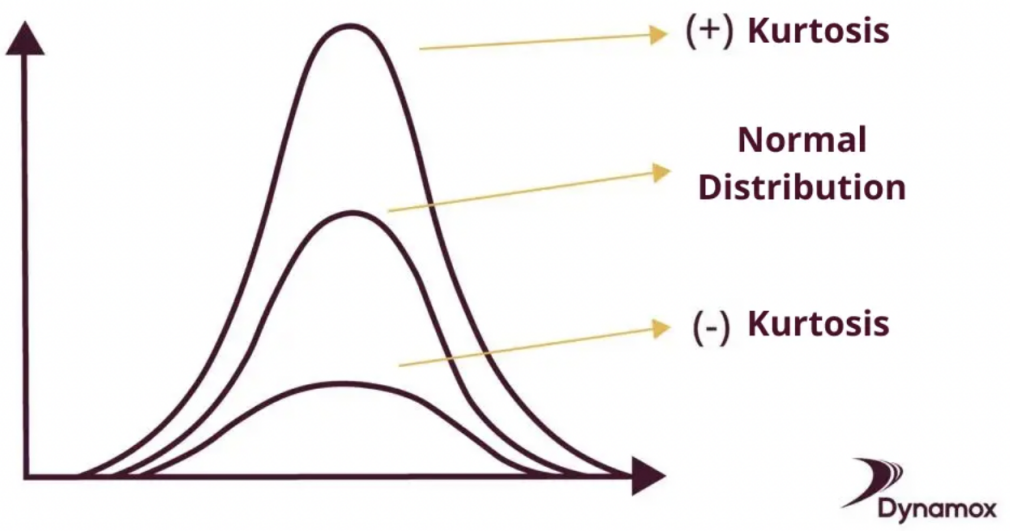
- **Uniform Distribution.**
- **Normal / Gaussian.**
- **Power Law - Exponential PDF.**
- **Binomial Probability Mass Function.**
- **Poisson Probability Mass Function.**

7.Percentiles and Moments

- **Percentiles**
  - A percentile shows the percentage of scores below a given value in a dataset.
- **Moments**
  - Quantitative measures of the shapes of a probability density function.
    - The First Moment is the Mean.
    - The Second Moment is the Variance.
    - The Third Moment is Skew.



    - The Fourth moment is Kurtosis.



8.A Crash Course in matplotlib

- **Line Graph**

```
plt.plot(x, norm.pdf(x))
plt.show()
```

- **Multiples Plots on a Graph**

```
plt.plot(x, norm.pdf(x, 1.0, 0.5))
```

- **Save it to a File**

```

plt.savefig('grafico.png', format='png')
```

- **Adjust the Axes**

```
axes = plt.axes()
axes.set_xlim([-5, 5])
axes.set_ylim([0, 1.0])
axes.set_xticks([-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5])
axes.set_yticks([0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8
```

- **Add a Grid**

```
axes.grid()
```

- **Change Line Types and Colors**

```
plt.plot(x, norm.pdf(x), 'b-') #aqui definindo linha e cor
plt.plot(x, norm.pdf(x, 1.0, 0.5), 'g:')
```

- **Labeling Axes and Adding a Legend**

```
plt.xlabel('Greebles')
plt.ylabel('Probability')
plt.legend(['Sneetches', 'Gacks'], loc=4)
```

- **XKCD Style**
- **Pie Chart**

```
plt.rcdefaults()
values = [12, 55, 4, 32, 14]
colors = ['r', 'g', 'b', 'c', 'm']
explode = [0, 0, 0.2, 0, 0]
labels = ['India', 'United States', 'Russia', 'China', 'Eu
plt.pie(values, colors= colors, labels=labels, explode = e
plt.title('Student Locations')
plt.show()
```

- **Bar Chart**

```
plt.bar(range(0,5), values, color= colors)
```

- **Scatter Plot**

```
from pylab import randn
X = randn(500)
Y = randn(500)
plt.scatter(X,Y)
plt.show()
```

- **Histogram Chart**

```
plt.hist(incomes, 50)
```

- **Box & Whisker Plot**
  - The line = Median. The box = 1st and 3rd quartiles. Half of the data = inside the box.
  - Good to identify Outliers

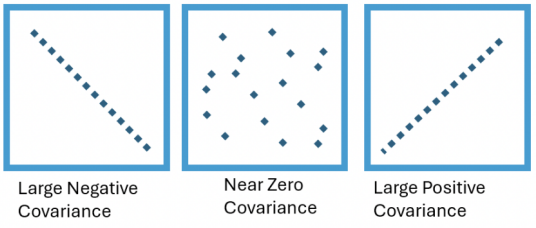
```
data = np.concatenate((uniformSkewed, high_outliers, low_o
plt.boxplot(data)
```

9.Advanced Visualization with Seaborn

- Notebook Practice.

10.Covariance and Correlation

- **Covariance**
  - Measures how two variables vary in tandem from their names.
  - Shows if two things change together (if one goes up, the other goes up or down). It doesn't tell how strong the change is, just if they change together.



- **Correlation**
  - Just divide the covariance by the standard deviations of both variables.
  - Use correlation to decide what experiments to conduct.
  - Shows how strongly and in what way two things are connected (if one goes up, the other goes up or down). It ranges from -1 to 1, where -1 is a perfect negative link, 1 is a perfect positive link, and 0 means no link.

11.Conditional Probability

- If I have two events that depend on each other, what's the probability that both will occur?
- A = 80% Passed the first test (easier). B= Passing the second test. 60% passed both tests.
  - What % who passed the first test also passed the second test?
  - 0.6/0.8 = 0.75.

```
from numpy import random
random.seed(0)

totals = {20:0, 30:0, 40:0, 50:0, 60:0, 70:0}
purchases = {20:0, 30:0, 40:0, 50:0, 60:0, 70:0}
totalPurchases = 0
for _ in range(100000):
    ageDecade = random.choice([20, 30, 40, 50, 60, 70])
    purchaseProbability = float(ageDecade) / 100.0 #Modify
    totals[ageDecade] += 1
    if (random.random() < purchaseProbability):
        totalPurchases += 1
        purchases[ageDecade] += 1
```

12.Bayes' Theorem

- Drug Testing Example:
  - Drug Test accurately identify user of drug 99% of the time.
  - Accurately has a negative result for 99% of non-users.
  - 0.3% of the overall population actually uses this drug.
- Drug Testing Solution
  - A = Is a user of the drug.
  - B = Tested positively for the drug.
  - P(B) = 1.3%(0.99 * 0.003 +0.01 * 0.997) = 22.8%

    - Someone being an actual user of the drug given that they tested positive is only 22.8%.
- Drug Testing Conclusion
    - Even though P(B|A) is 99%. It doesn't mean P(A|B) is high, in this case is really low.

Conclusões

In conclusion, this content covers the basics of different data types (numerical, categorical, ordinal) and key statistics like mean, median, and mode. It explains important concepts like variance, standard deviation, and probability distributions, and introduces visualization tools like matplotlib and seaborn. By exploring covariance, correlation, conditional probability, and Bayes' Theorem, it gives a clear and practical understanding of how to analyze and visualize data using Python.

Referências

Video: Data Science and Machine Learning. Available at https://www.udemy.com/course/data-science-and-machine-learning-with-python-hands-on/learn/lecture/15090172#overview. Accessed on: 09/13/2024.

Image: Covariance and Correlation Chart. Available at <//www.statisticshowto.com/probability-and-statistics/statistics-definitions/covariance/> Accessed on: 09/14/2024.

4o