Lecture 18 - Probability and Combinatorics Examples

DSC 40A, Winter 2024

Announcements

- Homework 6 is posted and due next Wednesday.
- HDSI undergrad & faculty mixer will be this afternoon 3-5pm at HDSI patio
  - Light refreshment will be provided

Agenda

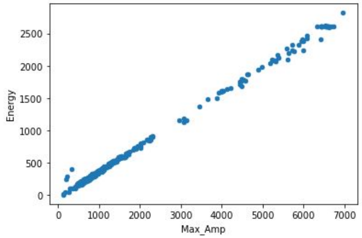
- Invited Algorithm Presentation
- Review of combinatorics.
- Lots of examples.

Invited Algorithm Presentation: Owen Shi

HW4 Algorithm

Owen Shi

In [5]:
```
waveforms.plot(kind='scatter', x='Max_Amp', y='Energy');
```

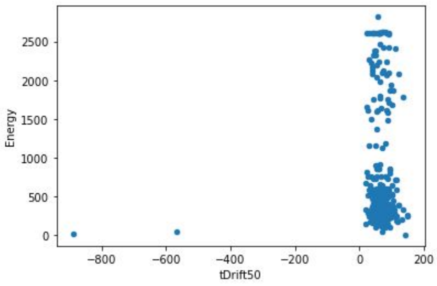


In [4]:
```
waveforms = pd.read_csv('HPGeData.csv')
waveforms.head()
```

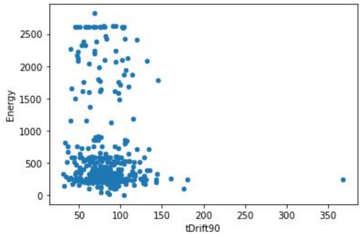
Out[4]:

	Max_Amp	tDrift50	tDrift90	tDrift100	blnoise	tslope
0	1233.0	61.0	69.0	81.0	11.5110	-108.1349
1	1319.0	93.0	116.0	135.0	3.7505	-112.9078
2	1237.0	81.0	89.0	104.0	2.3472	-111.2230
3	4469.0	90.0	98.0	112.0	1.8966	-111.4219
4	796.0	84.0	95.0	117.0	2.9256	-123.4542

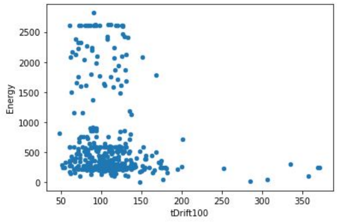
In [6]:
```
waveforms.plot(kind='scatter', x='tDrift50', y='Energy');
```



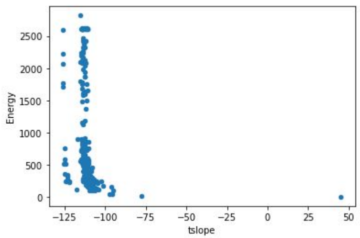
In [8]:
```
waveforms.plot(kind='scatter', x='tDrift90', y='Energy');
```



In [9]:
```
waveforms.plot(kind='scatter', x='tDrift100', y='Energy');
```



In [11]:
```
waveforms.plot(kind='scatter', x='tslope', y='Energy');
```



The Framework

```
In [5]: def design_matrix(d):
            df = d.copy()
            df['Intercept'] = 1
            feat1 = (1 / df['blnoise']**2 + 1 / df['tslope']) / df['Max_Amp']
            feat2 = (1 / df['blnoise'])**1.5 / (df['Max_Amp'] + df['blnoise']**3)
            feat3 = (df['Max_Amp'] * df['blnoise'] * df['tslope']) / \
                    (df['Max_Amp']**2 + 1 / df['blnoise'] + 1 / df['tslope'])
            df['feat1'] = feat1
            df['feat2'] = feat2
            df['feat3'] = feat3
        #     df.plot(kind='scatter', y='Energy', x='feat3')
            return df.get(['Intercept', 'Max_Amp', 'feat1', 'feat2', 'feat3']).to_numpy()

        design_matrix(waveforms)
```

```
In [6]: def observation_vector():
            return waveforms['Energy']
```

```
In [31]: np.random.seed(10)
         lambdas = np.logspace(-10, 10, 100)
         mses = []
         for l in lambdas:
             mse = 0
             w = w_star(l)
             for i in range(waveforms.shape[0]):
                 row = waveforms.iloc[i]
                 feat1 = 1 / row[4] + 1 / row[5]
                 feat2 = 1 / row[5] * 1 / feat1
                 X = pd.Series([1, row[0], feat1, feat2])
                 pred = X @ w
                 mse += (pred - row['Energy']) ** 2
             mse /= waveforms.shape[0]
             mses.append(mse)

         lambdas[np.argmin(mses)]

Out[31]: 4.132012400115335e-09

In [ ]: best_lambda = 0
```

```
[7]: def w_star(lam):
         X = design_matrix(waveforms)
         y = observation_vector()
         return np.linalg.inv(X.T @ X + lam * np.eye(X.shape[1])) @ X.T @ y
```

```python
def design_matrix(d):
    df = d.copy()
    df['Intercept'] = 1
    return df.get(['Intercept', 'Max_Amp']).to_numpy()
```

```python
def design_matrix(d):
    df = d.copy()
    df['Intercept'] = 1
    feat1 = 1 / df['blnoise']
    feat2 = 1 / df['tslope']
    df['feat1'] = feat1
    df['feat2'] = feat2
    return df.get(['Intercept', 'Max_Amp', 'feat1', 'feat2']).to_numpy()
```

```python
def design_matrix(d):
    df = d.copy()
    df['Intercept'] = 1
    feat1 = 1 / df['blnoise'] + 1 / df['tslope']
    feat2 = 1 / df['tslope'] * 1 / (1 / df['blnoise'] + 1 / df['tslope'])
    feat3 = df['tslope'] / df['Max_Amp']
    df['feat1'] = feat1
    df['feat2'] = feat2
    df['feat3'] = feat3
    return df.get(['Intercept', 'Max_Amp', 'feat1', 'feat2', 'feat3']).to_numpy()
```

```python
def design_matrix(d):
    df = d.copy()
    df['Intercept'] = 1
    feat1 = (1 / df['blnoise'] + 1 / df['tslope']) / df['Max_Amp']
    feat2 = df['Max_Amp'] / df['blnoise']
    feat3 = df['Max_Amp'] / df['tslope']
    df['feat1'] = feat1
    df['feat2'] = feat2
    df['feat3'] = feat3
    return df.get(['Intercept', 'Max_Amp', 'feat1', 'feat2', 'feat3']).to_numpy()
```

Submission History

#	Submitted On (PST)	Submitters	Score	Active
87	Feb 16 at 11:44 PM	OS	9.0	
86	Feb 16 at 11:43 PM	OS	0.0	
85	Feb 16 at 11:40 PM	OS	9.0	
84	Feb 16 at 10:01 PM	OS	9.0	
83	**Feb 16 at 9:58 PM**	OS	**9.0**	✔

```
without_f1 = 793.7921119778865
without_f2 = 766.2772193373274
without_f3 = 727.4278430354087
```

```python
In [5]: def design_matrix(d):
    df = d.copy()
    df['Intercept'] = 1
    feat1 = (1 / df['blnoise']**2 + 1 / df['tslope']) / df['Max_Amp']
    feat2 = (1 / df['blnoise'])**1.5 / (df['Max_Amp'] + df['blnoise']**3)
    feat3 = (df['Max_Amp'] * df['blnoise'] * df['tslope']) / \
            (df['Max_Amp']**2 + 1 / df['blnoise'] + 1 / df['tslope'])
    df['feat1'] = feat1
    df['feat2'] = feat2
    df['feat3'] = feat3
#     df.plot(kind='scatter', y='Energy', x='feat3')
    return df.get(['Intercept', 'Max_Amp', 'feat1', 'feat2', 'feat3']).to_numpy()

design_matrix(waveforms)
```

One More Extra Credit Opportunity

- Building a Naive Bayes classifier to separate neutrino signals from unwanted noises!
  - This one will be **Optional:** chances to earn extra credit, but does not count as part of homework problem.
  - Will be released and due together wit HW7
  - More details in the following weeks.
- The full HPGe dataset is released at https://zenodo.org/records/8257027
  - In raw waveform format, no extracted parameters.

Extra Credit Rules

- The classifier competition will earn you up to 10% extra credit on Midterm 2, depending on your leaderboard ranking
  - Same as the energy regression challenge
- However, the maximum extra credit you can earn from both challenges is capped at 10%
- Example: Owen ranked No. 2 on regression challenge, he will get 9% EC on Midterm 1, so the maximum amount of EC he can get on Midterm 2 is 1%
- This is to encourage students who did not get EC from the regression challenge to participate.

Review of combinatorics

Combinatorics as a tool for probability

- If S is a sample space consisting of equally-likely outcomes, and A is an event, then $P(A) = \frac{|A|}{|S|}$.
- In many examples, this will boil down to using permutations and/or combinations to count $|A|$ and $|S|$.
- **Tip:** Before starting a probability problem, always think about what the sample space S is!

Sequences

- A **sequence** of length *k* is obtained by selecting *k* elements from a group of *n* possible elements **with replacement**, such that **order matters**.

  True   True

- **Example:** You roll a die 10 times. How many different sequences of results are possible?

$6 \cdot 6 \cdot 6 \cdots\cdots 6$

1st roll   2nd roll   3rd roll

6^{10}

Sequences

In general, the number of ways to select k elements from a group of n possible elements such that **repetition is allowed** and **order matters** is

$$n^k.$$

Permutations

- A **permutation** is obtained by selecting *k* elements from a group of *n* possible elements **without replacement**, such that **order matters**.

  True   False

- **Example:** How many ways are there to select a president, vice president, and secretary from a group of 8 people?

$$\underset{\text{president}}{8} \cdot \underset{\text{VP}}{7} \cdot \underset{\text{Sect}}{6}$$

$$P(n,k) = P(8,3) = \frac{8!}{(8-3)!} = 8 \cdot 7 \cdot 6$$

Permutations

- In general, the number of ways to select k elements from a group of n possible elements such that **repetition is not allowed** and **order matters** is

$$P(n, k) = (n)(n-1)\ldots(n-k+1)$$

$$= \frac{n!}{(n-k)!}$$

Combinations

order doesnt matter

- A **combination** is a set of k items selected from a group of n possible elements **without replacement**, such that **order does not matter**.

False. False

- **Example:** How many ways are there to select a committee of 3 people from a group of 8 people?

$k=$ $n=$

$$\frac{P(8,3)}{3!} = C(8,3) = \binom{8}{3}$$

Choose a set of 3 people from 8

Combinations

In general, the number of ways to select k elements from a group of n elements such that **repetition is not allowed** and **order does not matter** is

False

False

$$C(n,k) = \binom{n}{k} = \frac{P(n,k)}{k!} = \frac{n!}{(n-k)!k!}$$

The symbol $\binom{n}{k}$ is pronounced "n choose k", and is also known as the **binomial coefficient**.

Lots of examples

Order matter? \ Replacement?	True	False
True	Sequence	Permutation
False	dominoes	combination

Sum of multiple combinations

replacemen? True
Order matter? False

Discussion Question

A domino consists of two faces, each with anywhere between 0 and 6 dots. A set of dominoes consists of every possible combination of dots on each face.

How many dominoes are in the set of dominoes?

a) $\binom{7}{2}$ → ① do not allow replacemen?

b) $\binom{7}{1} + \binom{7}{2}$

False False

c) $P(7,2)$

d) $\frac{P(7,2)}{P(7,1)}7!$

+ $\binom{7}{2}$

② Think about double dominoes?

$\binom{7}{1}$

Selecting students — overview

We're going answer the same question using several different techniques.

Question 1: There are 20 students in a class. Avi is one of them. Suppose we select 5 students in the class uniformly at random **without replacement**. What is the probability that Avi is among the 5 selected students?

Selecting students (Method 1: using permutations)

Question 1: There are 20 students in a class. Avi is one of them. Suppose we select 5 students in the class uniformly at random **without replacement**. What is the probability that Avi is among the 5 selected students?

S = a permutation (ordered selection) of 5 students chosen from A, B ... T

ex) LPFGA, LFGAT ....

P(A included) = # permutation w/ A / total # of permutations

$20 \cdot 19 \cdot 18 \cdot 17 \cdot 16 = P(20,5)$

Numerator: # of permutations including A.

ex). ACJTE

A _ _ _ _ → $1 \times 19 \times 18 \times 17 \times 16$

$1. \times 19 \times 18 \times 17 \times 16$

5 cases

A _ _ _ _

_ A _ _ _

_ _ A _ _

_ _ _ A _

_ _ _ _ A

total # of perm is

$5 \cdot 19 \cdot 18 \cdot 17 \cdot 16$

$= 5 \cdot P(19,4)$

$$\frac{5 \cdot P(19,4)}{P(20,5)} = \frac{5 \cdot \frac{19!}{15!}}{\left(\frac{20!}{15!}\right)} = 5 \cdot \frac{19!}{20!} = \frac{5}{20} = 1/4$$

Selecting students (Method 2: using permutations and the complement)

Question 1: There are 20 students in a class. Avi is one of them. Suppose we select 5 students in the class uniformly at random **without replacement**. What is the probability that Avi is among the 5 selected students?

$$\frac{\#\text{ Perms including A}}{\text{total \# Perms}} = \frac{\text{total \# of Perms} - \#\text{ Perms not include A}}{\text{total \# of Perms}}$$

$$\#\text{ perms not include A} = 19 \cdot 18 \cdot 17 \cdot 16 \cdot 15 = P(19,5)$$

$$\frac{P(20,5) - P(19,5)}{P(20,5)} = \frac{1}{4}$$

Selecting students (Method 3: using combinations)

Order = False

Question 1: There are 20 students in a class. Avi is one of them. Suppose we select 5 students in the class uniformly at random **without replacement**. What is the probability that Avi is among the 5 selected students?

S = set of 5 students chosen from A, B, ..., T

set → Order = False.

ex) {B, D, G, H, M}

$$P(A \text{ included}) = \frac{\#\text{ of sets of 5 students including A}}{\#\text{ of sets of 5 students}}$$

Selecting students (Method 3: using combinations)

Question 1, Part 1 (Denominator): If you draw a sample of size 5 at random without replacement from a population of size 20, how many different **sets** of individuals could you draw?

sets of 5 students: $C(20,5) = \binom{20}{5}$

$= \frac{20!}{15!\,5!}$

Selecting students (Method 3: using combinations)

Question 1, Part 2 (Numerator): If you draw a sample of size 5 at random without replacement from a population of size 20, how many different **sets** of individuals include Avi?

sets include Avi

$n = 19$

$k = 4$

$C(n,k) = C(19,4)$

# of other students except A B, C, ..., T

choose 4 other students to go w/ Avi

$P(\text{A included}) = \frac{C(19,4)}{C(20,5)} = \frac{1}{4}$

Selecting students (Method 3: using combinations)

Question 1: There are 20 students in a class. Avi is one of them. Suppose we select 5 students in the class uniformly at random **without replacement**. What is the probability that Avi is among the 5 selected students?

Selecting students (Method 4: “the easy way”)

Question 1: There are 20 students in a class. Avi is one of them. Suppose we select 5 students in the class uniformly at random **without replacement**. What is the probability that Avi is among the 5 selected students?

With vs. without replacement

Discussion Question

We've determined that a probability that a random sample of 5 students from a class of 20 **without replacement** contains Avi (one student in particular) is $\frac{1}{4}$.

Suppose we instead sampled **with replacement**. Would the resulting probability be equal to, greater than, or less than $\frac{1}{4}$?

a) Equal to
b) Greater than
c) Less than

Summary

Summary

- A **sequence** is obtained by selecting k elements from a group of n possible elements with replacement, such that order matters.
  - Number of sequences: n^k.

- A **permutation** is obtained by selecting k elements from a group of n possible elements without replacement, such that order matters.
  - Number of permutations: $P(n, k) = \frac{n!}{(n-k)!}$.

- A **combination** is obtained by selecting k elements from a group of n possible elements without replacement, such that order does not matter.
  - Number of combinations: $\binom{n}{k} = \frac{n!}{(n-k)!k!}$.