Wesleyan University, Fall 2021, COMP 411
Homework 2: Working With Probability
Due by 11:59pm on September 23, 2021

---

1. Basic Probability

Problem 1. *Work through examples in lecture slides, making sure you understand.*

Problem 2. *A box contains three marbles: one red, one green, and one blue. Consider an experiment that consists of taking one marble from the box then replacing it in the box and drawing a second marble from the box. What is the sample space? If, at all times, each marble in the box is equally likely to be selected, what is the probability of each point in the sample space?*

Problem 3. *Repeat Problem 1 when the second marble is drawn without replacing the first marble.*

Problem 4. *A coin is to be tossed until a head appears twice in a row. What is the sample space for this experiment? If the coin is fair, what is the probability that it will be tossed exactly four times?*

Problem 5. *Let E and F be mutually exclusive events in the sample space of an experiment. Suppose that the experiment is repeated until either event E or event F occurs. What does the sample space of this new super experiment look like? Show that the probability that event E occurs before event F is $P(E)/[P(E) + P(F)]$. Hint: Argue that the probability that the original experiment is performed n times and E appears on the nth time is $P(E)(1-p)^{n-1}, n = 1, 2, ...$, where $p = P(E) + P(F)$. Add these probabilities to get the desired answer.*

Problem 6. *The dice game craps is played as follows. The player throws two dice, and if the sum is seven or eleven, then she wins. If the sum is two, three, or twelve, then she loses. If the sum is anything else, then she continues throwing until she either throws that number again (in which case she wins) or she throws a seven (in which case she loses). Calculate the probability that the player wins.*

Problem 7. *Urn 1 contains two green balls and one black ball, while urn 2 contains one green ball and five black balls. One ball is drawn at random from urn 1 and placed in urn 2. A ball is then drawn from urn 2. It happens to be green. What is the probability that the transferred ball was green?*

2. Using Probability to Ask Questions About Your Data

Problem 8. *For two of the continuous valued features in your dataset from Homework 1, plot the values using a histogram, binning values.*

PROBLEM 9. *A univariate Gaussian distribution is a type of continuous random variable, with probability density function given by the following.*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where μ and σ are the parameters of the distribution. μ is the mean, σ is the standard deviation, and σ^2 is the variance. The variance is the average of the squared differences from the mean.

a: *For the features you used in the previous problem compute the the parameters, i.e., mean and standard deviation, for the univariate Gaussian distribution.*

b: *Download and import scipy. Use scipy to fit your (continuous) data for the features you used in the previous problem as follows.*

```
import numpy as np
from scipy.stats import norm
import matplotlib.pyplot as plt
data = np.random.normal(loc=5.0, scale=2.0, size=1000)
mean,std=norm.fit(data)
```

The mean and standard deviation returned should match what you had in the previous part. See `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.norm.html`

c: *The normal distribution may not however be the best distribution for your data. What can we do? We can try fitting to other distributions and then computing some statistics to evaluate goodness of fit.*

```
list_of_dists = ['alpha','anglit','arcsine','beta','betaprime','bradford',\
'burr','burr12','cauchy','chi','chi2','cosine','dgamma','dweibull',\
'erlang','expon','exponnorm','exponweib','exponpow','f','fatiguelife',\
'fisk','foldcauchy','foldnorm','frechet_r','frechet_l','genlogistic',\
'genpareto','gennorm','genexpon','genextreme','gausshyper','gamma',\
'gengamma','genhalflogistic','gilbrat','gompertz','gumbel_r','gumbel_l',\
'halfcauchy','halflogistic','halfnorm','halfgennorm','hypsecant',\
'invgamma','invgauss','invweibull','johnsonsb','johnsonsu','kstwobign',\
'laplace','levy','levy_l','logistic','loggamma','loglaplace','lognorm',\
'lomax','maxwell','mielke','nakagami','ncx2','ncf','nct','norm','pareto',\
'pearson3','powerlaw','powerlognorm','powernorm','rdist','reciprocal',\
'rayleigh','rice','recipinvgauss','semicircular','t','triang','truncexpon',\
'truncnorm','tukeylambda','uniform','vonmises','vonmises_line','wald',\
'weibull_min','weibull_max']

results = []
for i in list_of_dists:
    dist = getattr(stats, i)
    param = dist.fit(data)
    a = stats.kstest(data, i, args=param)
    results.append((i,a[0],a[1]))
    results.sort(key=lambda x:float(x[2]), reverse=True)
for j in results:
    print("{}: statistic={}, pvalue={}".format(j[0], j[1], j[2]))
```

Read up on the KS-test and how you should use it to interpret goodness of fit of your data to each distribution. You may have to do some google searching.
`https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kstest.html`

d: *Let's go back to the normal distribution we fitted. Becasue it is continuous, we can't ask questions about individual values using the probability density function (probability will always be 0). But we can ask questions using the cumulative distribution function. There is, however, no closed form for the cdf of the normal distribution, though an approximation can be had using the standard normal distribution (see*
`https://mathworld.wolfram.com/NormalDistributionFunction.html`*). Instead, what we will do is use scipy functions to do the cdf integration needed for the normal distribution cdf for us:*

```
norm(loc = mean, scale = std).cdf(x)
```

This will give us the probability that the variable X generating the data for the feature takes a value less than or equal to x. More here:
`https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.norm.html` *For the largest and smallest values in each of your feature vectors, l and s respectively, plug them into the CDF equation for the normal distribution to compute the probability that the random variable X generating the values for a feature has values less than or equal to l or s. Try out a few other values from your features too.*

3. Feedback on slides

Let me know what you found easy and hard to understand on the slides. Any comments or criticisims are appreciated.