MATH 728D: Machine Learning Lab #4: Probability

John Burkardt

December 5, 2018

Two flies land at random points in the unit circle. On average, what is the distance between them?

Probability and Statistics predict and analyze the behavior of systems subject to uncertainty or randomness. Of particular interest are

- the construction of distribution functions beforehand;
- the construction of histograms of observed events;
- the average and variance of predicted and observed events;
- the sampling or simulation of random processes;

1 Uniform Random Points in a Circle

To simulate the problem, we need to select points from the circle in a uniform random way. That means that the probability of drawing from any particular region is proportional to its area.

We need a procedure *ran(a,b)* that samples uniformly from interval (a, b). In MATLAB, this can be done by

```
r = a + (b-a) * rand ( );
```

and in the common case where $a = 0$, we simply have

```
r = b * rand ( );
```

Let's evaluate these proposed sampling methods for finding points (r, θ) or (x, y) in the unit circle:

1. $r = ran(0, 1)$, $\theta = ran(0, 2\pi)$;
2. $r = \sqrt{ran(0, 1)}$, $\theta = ran(0, 2\pi)$;
3. $x = ran(-1, 1)$, $y = ran(-1, 1)$, reject if $x^2 + y^2 > 1$.

Exercise 1:

1. Generate and plot 1000 points in the unit circle for method #1;
2. Generate and plot 1000 points in the unit circle for method #2;
3. Generate and plot 1000 points in the unit circle for method #3;
4. Which methods seem to sample the circle uniformly?

2 Average Distance in the Circle

We wish to do a simulation that computes an estimate `lbar` for the average distance $\bar{\ell}$ between a pair of uniformly random points in the unit circle. We also want an estimate `lvar` for the variance, σ^2, defined by

$$\sigma^2 = \sum_{i=1}^{n} \frac{\ell_i - \bar{\ell}}{n-1}$$

Exercise 2: Write a program that does the following:

- Generate 1000 random pairs of points;
- Determine the pairwise distances `l(i)`.
- Compute `lbar = mean ( l );`
- Compute `lvar = var ( l );`

Theoretically, the mean value should be $\bar{\ell} = \frac{128}{45\pi}$ and the variance should be $\sigma^2 = \frac{2025\pi^2 - 128^2}{45^2\pi^2}$. Compare your results to these values.

3 Distribution of Distances

Let us write $p(\ell)$ to represent the probability of observing the distance ℓ between a pair of randomly chosen points in the unit circle. We know then that $p(\ell)$ is nonzero only for $0 \leq \ell \leq 2$, and that $\int_0^2 p(\ell) d\ell = 1$, and we have seen that $\bar{\ell}$ is near 1. To get a better feeling for the variation of ℓ, we can construct a histogram.

Exercise 3:

- Use circle sampling method #2 to generate 10,000 pairs of points;
- Compute the distances `l(i)`.
- Use the MATLAB command **histogram()** to histogram the data;
- Modify the **histogram()** command to use the pdf normalization:

```
histogram ( l, 'Normalization', 'pdf' )
```

- Issue the command `hold on` so you can add to this plot;
- Set `d=linspace ( 0.0, 2.0, 101)`, then evaluate

$$p = \frac{1}{\pi} d \left(4 \arccos(d/2) - d\sqrt{4 - d^2}\right)$$

  and issue the command

```
plot ( d, p )
```

- The pdf curve should match the shape of the histogram;

4 A Classification Experiment

Factory #1 makes candy with an average weight of $\mu_1 = 10$ ounces, with a standard deviation $\sigma_1 = 1$ ounce. For Factory #2, the statistics are $\mu_2 = 13$ and $\sigma_2 = 3$. A mixture of candy has arrived, and it is necessary to try to estimate which factory each piece of candy came from. For the PDF's, use the normal distribution $N(\mu, \sigma^2)$.

Exercise 4:

1. Look at a plot of the two PDF's:
   - On a single plot, display pdf_1 and pdf_2 over the range $0 <= x <= 25$;
   - Estimate a value x where the PDF's are equal;
2. Generate the data:
   - Using the function `normal_samples(n,mu,sigma)`, generate 1000 samples of pdf_1 as `x1`, and 500 samples of pdf_2 as `x2`;
   - Display the sample overlap with these commands:

```
h1 = histcounts ( x1, 0:25 );
h2 = histcounts ( x2, 0:25 );
bar ( [h1',h2'], 'stacked' );
```

3. Attempt to classify the data
   - Concatenate `x1` and `x2` into a single array `x`;
   - For each `x(i)`, evaluate pdf_1 and pdf_2.
   - Assign `x(i)` to class 1 if $pdf_1 > pdf_2$, otherwise to class 2;
   - Count `correct`, the number of correct assignments;
4. Evaluate the classification and print the score:

```
score = correct / 1500;
```

You may notice that some values less than 10 get assigned to class 2. Can you explain why this happens?