Homework 04

Brown University

DATA 1010

Fall 2019

Problem 1

Suppose that the probability density function for the random drawing
point where your dart hits the dartboard $D \subset \mathbb{R}^2$ is given by $$f(x,y) = \frac{1}{\pi} e^{-x^2 - y^2},$$ where the origin is situated at the dartboard's bull's eye, and where $x$ and $y$ are measured in inches (this function is positive everywhere in $\mathbb{R}^2$, so the "dartboard" includes the disk shown as well as the (infinite) wall it is mounted on—this is realistic insofar as one can indeed hit the wall with a dart throw).
Find the probability of scoring triple 20 on your next throw.
(Optional) Confirm your result using Monte-Carlo simulation.

Note: The triple 20 region is the smaller of the two thin red strips in the sector labeled "20". The inner and outer radii of this thin strip are 3.85 inches and 4.2 inches, respectively.

solution

The region in question is described most easily in polar coordinates: it is the set of points whose polar coordinates (r, θ) satisfy $r_i \le r \le r_o$ and* $81^\circ \le \theta \le 99^\circ$, where $r_i = 3.85$ and $r_o = 4.2$. (Note that the width of each sector is $360^\circ/20 = 18^\circ$, so the angles of the rays bounding the sector labeled 20 are $90^\circ \pm \frac{18^\circ}{2}$)

Therefore, we can obtain the probability of hitting the triple 20 by expressing the density function in polar coordinates and integrating:

$$\frac{1}{\pi} e^{-r^2} r\, dr\, d\theta = \left(\frac{\pi}{10}\right) \left(\frac{1}{\pi}\right) \left(-\tfrac{1}{2} e^{-r_o^2} - \left(-\tfrac{1}{2} e^{-r_i^2}\right)\right) . \$\$ Substituting the given value$$

Problem 2

Suppose that X and Y have joint PDF $f(x, y) = \frac{3}{2}y$ on the upper unit disk (that is, the set of points which have positive y-coordinate and are less than one unit from the origin).

1. Verify that f is indeed a probability density function.
2. Find the density of the distribution of X.
3. Find the conditional density of Y given $X = x$.
4. Find $\mathbb{E}[Y|X]$.

solution

1. We have $\int_0^\pi \int_0^1 \frac{3}{2}r\sin\theta(r\,dr\,d\theta) = 1$, so f is indeed a probability density function.
2. The distribution of X is obtained by integrating out y:

```
\sqrt{1-x^2}} \, d y = \int_{0}^{\sqrt{1-x^2}} \frac{3}{2}y
    \, d y = \frac{3}{4}(1-x^2).$$
```

3. The conditional density of Y given $X = x$ is the joint density divided by X's marginal density at x:

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x, y)}{\frac{3}{4}(1-x^2)} = \frac{\frac{3}{2}\boldsymbol{y}\mathbf{1}_{0\leq y\leq\sqrt{1-x^2}}}{\frac{3}{4}(1-x^2)} = \frac{\boldsymbol{2y}\mathbf{1}_{0\leq y\leq\sqrt{1-x^2}}}{1-x^2}.$$

4. The conditional expectation of Y given X is obtained by integrating y times the conditional density of Y given $X = x$ and then substituting X for x:

```
\sqrt{1-x^2}}}{1-x^2} \, d y=
    \int_{0}^{\sqrt{1-x^2}}\frac{2y^2}{1-x^2} \, d y =
    \frac{2(1-x^2)^{3/2}}{3(1-x^2)} = \frac{2}{3}\sqrt{1-x^2}.$$
```

Problem 3

The *skewness* of a distribution ν is a measure of its asymmetry about its mean. It is defined to be

$$\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right],$$

where X is a random variable with distribution ν, μ is the mean of X, and σ is the standard deviation of X. Find the skewness of the exponential distribution with parameter 1. You should set up the integrals on your own, but feel free to evaluate them using a symbolic computation system.

solution

We begin by calculating

$$1/\lambda, \$\$ and$$

```
2/\lambda^2,$$ so $\sigma = \sqrt{\mathbb{E}[X^2] - \mathbb{E}[X]^2} = 1/\la
```

Finally,

$$\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \int_0^\infty \frac{(x-\mu)^3}{\sigma^3}\lambda e^{-\lambda x}\, dx = 2$$

Evaluating these integrals using `SymPy`:

```
using SymPy
@vars x λ
integrate(x*λ*exp(-λ*x),(x,0,oo))
integrate(x^2*λ*exp(-λ*x),(x,0,oo))
integrate((x-1/λ)^3/(1/λ)^3*λ*exp(-λ*x),(x,0,oo))
```

Problem 4

Let X be the first digit of the number of residents of a randomly selected world city. What would you expect the distribution of X to look like? What about the *last* digit Y?

Load the associated world city populations CSV as a DataFrame and check your predictions. Compare to the distribution with probability mass function

$$m(d) = \log_{10}(d+1) - \log_{10}(d) \quad \text{for } d \in \{1, 2, \ldots, 9\}.$$

```
using StatsBase, Plots, FileIO, DataFrames
D = DataFrame(load("cities.csv"))
D[:Population]
tallydict = # you fill in this part
sticks(1:9,collect(values(tallydict)))
```

solution

One natural prediction is that both distributions should be uniform (on 1:9 for the first digit, 0:9 for the last). Let’s investigate.

```
using StatsBase, Plots, FileIO, DataFrames
D = DataFrame(load("cities.csv"))
D[:Population]
tallydict = sort(countmap([string(n)[1] for n in D
ys = collect(values(tallydict))
sticks(0:9,ys/sum(ys),label="first digit proportio
sticks!((1:9).+0.1,[1/9 for d=1:9],label="uniform"
```

drawing

This not seem like a good fit. Let’s try the distribution suggested in the problem statement.

```
sticks(1:9,ys/sum(ys),label="first digit proportio
sticks!((1:9).+0.1,[log10(d+1)-log10(d) for d=1:9]
```

drawing

This fit seems better. This distribution is called Benford’s distribution, and it fits a variety of real-world leading-digit data, for reasons that are not completely well understood.

Now, let's plot the distribution of the last digit:

```
tallydict = sort(countmap([string(n)[end] for n in
ys = collect(values(tallydict))
sticks(0:9,ys/sum(ys),label="last digit proportion
sticks!((0:9).+0.1,[1/10 for d=0:9],label="uniform
```

drawing

We can see that our prediction was pretty accurate for the last digit, except that this data set contains quite a few rounded numbers which causes 0 to be overrepresented.

Problem 5

A **call option** is a financial contract between two parties which grants the buyer the right, but not the obligation, to purchase a specified security at a specified price (called the **strike price**) at a specified date in the future (called the **expiration date**).

Suppose that you purchase a call option for 10 shares of AAPL with a strike price of $\$216$ and an expiration 22 business days from now. Suppose that the daily change in the price of AAPL is normally distributed with mean zero and standard deviation $\$8.44$, and that the changes for different days are independent.

(a) Find a function f such that the call option is worth $f(P)$ dollars to you if the share price in 22 days is P. Draw a graph of f. Hint: if the price is greater than $\$216$, would you exercise the option and purchase the stock? What if it's less than $\$216$?

(b) Find the distribution of P.

(c) Find the fair price of the call option, based on the above assumptions.

Notes: (1) the data in this problem are real: the current price at time of writing is $\$216$, and the daily fluctuations have had an empirical standard deviation of $\$8.44$ historically. The number of business days in a month is approximately 22. (2) Although this problem uses finance ideas, all of the finance information you need to solve the problem is in the problem statement.

To help you with the symbolic integration, here's some code for finding the expected value of the Gaussian distribution centered at μ.

[2]

```
using SymPy
@vars P σ μ positive=true
I = integrate(P*1/sqrt(2*PI*σ^2)*exp(-(P-μ)^2/(2σ^2)),(P,-oo,oo))
```

$$\mu$$

solution

(a) We have $f(P) = \max(0, P - 216)$, since if the price is in excess of $\$216$, we can sell it and make a profit equal to the difference between the price and $\$216$. If the price is less, we would not exercise the option and it would be worthless.

(b) The distribution of the price of the stock in 22 days is Gaussian with mean 216 and variance $\$8.44 \cdot 22 = \$185.68$.

(c) The expected value of the option is

$$\int_{-\infty}^{\infty} f(P)\phi(P)\, dP,$$

where $\phi(P)$ is the Gaussian density with mean 0 and variance 185.68. The code block

```julia
using SymPy
@vars P σ μ positive=true
I = integrate((P-μ)*1/sqrt(2*PI*σ^2)*exp(-(P-μ)^2/(2σ^2)),P,μ,oo)
```

returns $\frac{\sigma}{\sqrt{2\pi}}$, so we can say that the fair price of the option is $\frac{\sqrt{185.68}}{\sqrt{2\pi}} \approx \$5.44$ dollars. (Note that we could alternatively evaluate this integral by hand using substitution).

Problem 6

Consider two random variables X and Y whose joint distribution has probability mass of $\frac{1}{n}$ at each of the n points $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ in $\mathbb{R}^2$. Show that the covariance matrix of X and Y is equal to

$$\frac{1}{n} \sum_{i=1}^{n} \left[\begin{smallmatrix} x_i - \overline{x} \\ y_i - \overline{y} \end{smallmatrix}\right] \left[\begin{smallmatrix} x_i - \overline{x} & y_i - \overline{y} \end{smallmatrix}\right].$$

where $\overline{x} = (x_1 + \cdots + x_n)/n$ and $\overline{y} = (y_1 + \cdots + y_n)/n$.

solution

The off-diagonal entries of the covariance matrix are each equal to

$$\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

where μ_X and μ_Y are the expected values of X and Y. Then using the formula $\mathbb{E}[g(X, Y)] = \sum_{(x,y) \in \mathbb{R}^2} g(x, y) m_{X,Y}(x, y)$, we find that

$$\mathbb{E}[XY] = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}),$$

Meanwhile, the off-diagonal entry of $\frac{1}{n} \sum_{i=1}^{n} [x_i - \overline{x}, y_i - \overline{y}][x_i - \overline{x}, y_i - \overline{y}]'$ is

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}),$$

which is indeed equal to the expression we found for $\mathbb{E}[XY]$. Similar analysis applies to the diagonal entries.

Problem 8

The *Epanechnikov* kernel is defined by

$$D(u) = \frac{3}{4}(1 - u^2)\mathbf{1}_{|u|\leq 1}.$$

- Is D continuous? Is it differentiable? Is it twice differentiable?
- Is the tri-cube weight function continuous? Is it differentiable? Is it twice differentiable?

Feel free to use SymPy to perform the symbolic differentiation in this problem.

solution

We calculate derivatives of the polynomial expressions in the two functions:

```
using SymPy
@vars u
subs(diff(1-u^2,u),u=>1)
subs(diff((1-u^3)^3,u),u=>1)
subs(diff((1-u^3)^3,(u,2)),u=>1)
```

We find that the Epanechnikov kernel is continuous but not differentiable at 1, since its derivative from the right is zero and its derivative from the left is negative. The tri-cube weight function is twice differentiable, since its first and second derivatives from the left are zero at 1.

Problem 9

Simulate $n = 1000$ samples from the joint distribution of X and Y, given that X is uniform on $[0, 1]$ and $Y = 2 + 1.2X + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.5)$. Record the integrated squared error for the Nadaraya-Watson estimator (with bandwidth selected by cross-validation) and for the line of best fit.

Notes: (1) You can approximate the integrated squared difference between two functions by evaluating the squared difference at the points of a fine-mesh grid. And (2) you'll have to write code for simulating from the joint distribution of X and Y, but then the Nadara-Watson part you can mostly get from Data Gymnasia.

solution

We begin by loading the optimization package, defining the regression function, and defining a function to draw samples from the given distribution.

```julia
# solution
using Optim
r(x) = 2 + 1.2x
function sampleXY()
    X = rand()
    Y = r(X) + sqrt(0.5)*randn()
    (X,Y)
end

n = 1000
samples = [sampleXY() for i=1:n]
xs = 0:1/2^8:1
ys = 0:1/2^5:6
```

solution

Next we do kernel density estimation with cross validation.

```julia
# solution
D(u) = abs(u) < 1 ? 70/81*(1-abs(u)^3)^3 : 0 # tri-cube function
D(λ,u) = 1/λ*D(u/λ) # scaled tri-cube
K(λ,x,y) = D(λ,x) * D(λ,y) # kernel
```

```
kde(λ,x,y,samples) = sum(K(λ,x-Xi,y-Yi) for (Xi,Yi) in samples)/length(samples)

function kdeCV(λ,i,samples)
    x,y = samples[i]
    newsamples = copy(samples)
    deleteat!(newsamples,i)
    kde(λ,x,y,newsamples)
end

# first line approximates ∫f̂², the second line approximates -(2/n)∫f̂f
J(λ) = sum([kde(λ,x,y,samples)^2 for x=xs,y=ys])*step(xs)*step(ys) -
     2/length(samples)*sum(kdeCV(λ,i,samples) for i=1:length(samples))
λ_best_cv = optimize(λ->J(first(λ)),[1.0],BFGS()).minimizer[1]
r̂(λ,x) = sum(D(λ,x-Xi)*Yi for (Xi,Yi) in samples)/sum(D(λ,x-Xi) for (Xi,Yi) in samples)
```

solution

Finally, we approximate the integrated squared error for the nonparametric method as well as for the parametric method.

```
# solution
ISE_nonparametric = sum((r̂(λ_best_cv,x) - r(x))^2 for x in xs)

X = [ones(length(samples)) [x for (x,y) in samples]]
β = (X'*X) \ X' * [y for (x,y) in samples]
ISE_linear = sum((β⋅[1,x]-r(x))^2 for x in xs)
```

solution

We find that the integrated squared error is much lower for the linear approximation, which makes sense because the regression function is in fact linear. In other words, the inductive bias of the model aligns well with actual probability measure, and that leads to increased accuracy relative to a model with less inductive bias.

Problem 10

- Find the variance of the uniform distribution on the interval $[0, 10]$.
- Generate 10 independent samples from the uniform distribution, calculate the average $\overline{X}$ for those samples, and estimate the variance as $\widehat{V} = \frac{1}{10}\sum_{i=1}^{10}(X_i - \overline{X})^2$. Package this whole process as a function, and call it a million times to find the mean of $\widehat{V}$.
- Which is larger, the answer to (a) or the answer to (b)? Calculate the percent error.

solution

We package the sampling procedure as a function and call it M times for $M = 10^6$:

```
function variance_estimate(n)
    X = [10*rand() for i=1:n]
    X̄ = mean(X)
    1/n * sum((x - X̄)^2 for x in X)
end
M = 10^6
variance_estimate_mean = mean(variance_estimate(10) for i=1:M)
true_variance = 10^2 / 12
perc_error = (variance_estimate_mean - true_variance)/true_variance
```

We find that the variance estimate is lower than the true variance, with approximately 10% error. If we repeat for other values of M, we find that this percent error is not diminishing.

```
[ ]
```

Problem 11

The Student's *t*-distribution with parameter ν is the distribution of the random variable

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$$

where $n = \nu + 1$, where $X_1, \ldots, Xn$ is a sequence of independent $N(\mu, \sigma^2)$'s, where $\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$, and where $S_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$

Estimate the variance of the Student's *t*-distribution with parameter $\nu = 10$ by using the above description to sample from it M times for some large M. Then compute the variance of the distribution which places probability mass $1/M$ at each of the simulated samples.

Look up the exact formula for the variance of the Student's *t*-distribution on Wikipedia and check that your result is close to the true value.

solution

We write a function which samples from the given distribution a large number of times. For convenience, we take $\mu = 0, \sigma = 1$.

```
function sampleT(v=10)
    n = v + 1
    X = randn(n)
    X̄ = mean(X)
    S = √(sum((x - X̄ )^2 for x in X)/(n-1))
    X̄ /(S/√(n))
end
M = 10^6
samples = [sampleT() for i=1:M]
m = mean(samples)
sum((s - m)^2 for s in samples)/M
```

The formula we discover on Wikipedia is $\nu/(\nu - 2)$, which is very close to the value obtained by our Monte Carlo simulation (approximately 1.25).