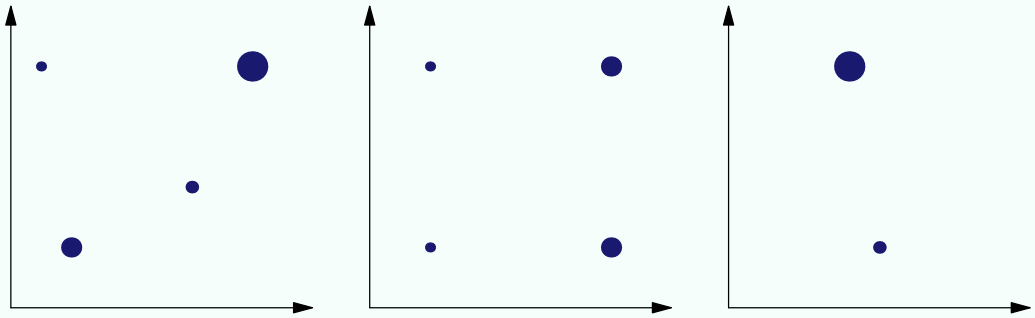
DATA 1010
Problem Set 8
Due 02 November 2018 at 11 PM

Problem 1

Each of the following figures shows the PMF of the joint distribution of two random variables (where point size is reflective of probability mass).



(a) For each figure, indicate whether the two random variables are positively correlated, negatively correlated, or uncorrelated.

(b) In which figure are the random variables independent?

(c) In the second figure, which random variable is uniformly distributed on its support?

(d) Estimate the mean of the random vector $[X, Y]$ for the third distribution (assuming that the two random variables are X and Y). Express your answer by drawing on the figure.

Solution

(a) Positively correlated, uncorrelated, negatively correlated. To check whether the covariance is positive, negative, or zero, we identify the center of mass of the arrangement of masses, and we count masses northeast or southwest of that point as positive, and ones northwest or southeast as negative. Each such contribution is proportional to the amount of mass at the point and to the distance from the vertical and horizontal lines through the center of mass.

(b) Only in the second figure. We can see that the PMF factors as the product of its marginals (both of which have positive mass at two real numbers).

(c) The second random variable (represented by the y-axis) is uniform, since the total mass along each of the two horizontal lines is the same.

(d) The center of mass is on the line segment connecting the two points, considerably closer to the larger mass.

Problem 2

Suppose that X is chosen uniformly at random from the interval $[0, 1]$, and then Y is chosen uniformly at random from the interval $[0, X]$.

(a) The information above specifies $f_{Y \mid X=x}(y)$, the conditional density of Y given that $X = x$. Find $f_{Y \mid X=x}(y)$.

(b) Use (a) to find the joint density $f_{X,Y}(x, y)$ of X and Y.

(c) Use (b) to find the marginal distribution of Y.

(d) You should find that the density of Y is *unbounded*. Explain why it isn't a contradiction for a probability density function to be unbounded (considering that the total amount of probability mass must be 1).

Solution

(a) The conditional density is uniform on $[0, x]$, so $f_{Y \mid X=x}(y) = \frac{1}{x}\mathbf{1}_{y \in [0,x]}$.

(b) The joint density function is the product of the marginal of X and the conditional density of Y with respect to X:

$$f_{X,Y}(x, y) = f_{Y \mid X=x}(y) f_X(x) = \frac{1}{x}\mathbf{1}_{0 \le y \le x \le 1}.$$

(c) Integrating, we find that

$$f_Y(y) = \int_{-\infty}^{\infty} \frac{1}{x}\mathbf{1}_{0 \le y \le x \le 1}\, \mathrm{d}x = \int_y^1 \frac{1}{x}\, \mathrm{d}x = \ln(1/y).$$

(d) There is no contradiction because $f_Y(y)$ represents probability *density*, not probability itself. If a fairly small but positive amount of mass is crammed into a very tiny region, then the density is very large (even though the total mass in that region is small).

Problem 3

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

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

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

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(ν=10)
    n = ν + 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).

Problem 4

Recall the probability mass function struct `PMF` we defined in Problem Set 7. Define a sampling method for it. Your function should return a random value whose distribution is the one represented by the given PMF.

```
function sample(P::PMF)
# ...
end
```

Solution

The idea is to use inverse CDF sampling: we split up the unit interval into segments whose lengths are equal to the masses of the PMF, we determine which interval a uniform random number falls in, and we return the corresponding value:

```
struct PMF
    masses
    values
end

function sample(P::PMF)
    markers = cumsum(P.masses)
    @assert markers[end] ≈ 1.0
    U = rand()
    for i = 1:length(markers)
        if U < markers[i]
            return P.values[i]
        end
    end
end
P = PMF([1/3,1/3,1/3],[0,1,2])
```

Problem 5

A random walker begins at one vertex of a square, and it repeatedly moves along one of the adjacent edges (chosen uniformly at random) to another vertex. Find the distribution of the number of steps N when the walker first reaches the vertex diagonally opposite to the starting vertex.

Solution

On every even-numbered step, the walker has a 50% chance of reaching the destination vertex and a 50% chance of heading back to the starting vertex. Therefore, the distribution has probability mass function

$$m(k) = (1/2)^{k/2} \quad \text{if } k \in \{2, 4, 6, \ldots\}.$$

Problem 6

In this problem we will show that a sum Z of independent standard normal random variables X and Y is normal with mean zero and variance 2.

(a) Let F_Z be the CDF of Z. Express $F_Z(z)$ as an integral over a subset of $\mathbb{R}^2$.

(b) You should find that the integrand in your answer to (a) is rotationally symmetric. Rotate the region of integration so that its boundary is a vertical line.

(c) Simplify the integral you found in (b) and show that it is equal to $\Phi_{0,2}(z)$, where Φ_{μ,σ^2} denotes the CDF of $\mathcal{N}(\mu, \sigma^2)$.

Note: this idea can be extended to show that an $\mathcal{N}(\mu_1, \sigma_1^2)$ plus an independent $\mathcal{N}(\mu_2, \sigma_2^2)$ has distribution $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Solution

(a) We have

$$F_Z(z) = \mathbb{P}(X + Y \leq z) = \iint_{x+y\leq z} \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}\, dx\, dy.$$

(b) The integrand is indeed rotationally symmetric (about the origin), so we can apply a rotation which maps the line $x + y = z$ to a vertical line. Such a rotation preserves the line's distance to the origin, which is $z/\sqrt{2}$. Therefore, the desired integral equals

$$\iint_{x\leq z/\sqrt{2}} \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}\, dx\, dy.$$

(c) We can perform this integral by first integrating with respect to y. Since $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1$, the above integral simplifies to

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z/\sqrt{2}} e^{-x^2/2}\, dx.$$

We can perform a u-substitution with $u = z/\sqrt{2}$ in this integral to get

$$\frac{1}{\sqrt{2\pi \cdot 4}} \int_{-\infty}^{z} e^{-u^2/4}\, du,$$

which is equal to $\Phi_{0,2}(z)$, as desired.

Problem 7

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.

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

```
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 $\sqrt{185.68}/\sqrt{2\pi} \approx \$5.44$ dollars. (Note that we could alternatively evaluate this integral by hand using u-substitution).

Problem 8

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: you can approximate the integrated squared difference between two functions by evaluating the squared difference at the points of a fine-mesh grid.

Solution

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

```
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
```

Next we do kernel density estimation with cross validation.

```
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 ∫f^2, the second line approximates -(2/n)∫f^f
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]
f̂(λ,x) = sum(D(λ,x-Xi)*Yi for (Xi,Yi) in samples)/sum(D(λ,x-Xi) for (Xi,Yi) in samples)
```

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

```
ISE_nonparametric = sum((f̂(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)
```

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.