Brown University
DATA 1010
Fall 2018: Final Exam
Samuel S. Watson

Name:

You will have three hours to complete the exam, which consists of 40 questions. Among the first 36 questions, you should only solve problems for standards for which you want to improve your medal from the second exam.

No calculators or other materials are allowed, except the provided reference sheets.

You are responsible for explaining your answer to **every** *question. Your explanations do not have to be any longer than necessary to convince the reader that your answer is correct.*

For questions with a final answer box, please write your answer as clearly as possible and strictly in accordance with the format specified in the problem statement. Do not write anything else in the answer box. Your answers will be grouped by Gradescope's AI, so following these instructions will make the grading process much smoother.

I verify that I have read the instructions and will abide by the rules of the exam: _____

Problem 1 [SETFUN]

Consider a function f from a set A to a set B. Suppose that $V \subset B$ has the property that $f^{-1}(B) = \emptyset$. Find $f^{-1}(B \setminus V)$.

Solution

Since no points in A map to any points in V, every point in A maps to some point in $B \setminus V$. Therefore, $f^{-1}(B \setminus V) = A$.

Final answer:

$$A$$

Problem 2 [JULIA]

Write a Julia function which removes the vowels from a word (treating y as a vowel). Your function should pass the test

```
@assert remove_vowels("grasshopper") == "grsshppr"
```

Solution

We filter the characters in the string using an array comprehension:

```
function remove_vowels(S)
    join([c for c in S if !(c in "aeiouy")])
end
@assert remove_vowels("grasshopper") == "grsshppr"
```

Problem 3 **[LINALG]**

Consider a list of seven nonzero vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_7$, and suppose that

$$\mathbf{v}_1 + \mathbf{v}_2 + \mathbf{v}_3 = \mathbf{0}.$$

Suppose further that this list has a linearly independent sublist of length 6. Is it necessarily the case that $\{\mathbf{v}_4, \mathbf{v}_5, \mathbf{v}_6\}$ is linearly independent?

Solution

Any linearly independent sublist of length 6 must exclude one of the first three vectors, because of the given linear dependence relation. Since $\{\mathbf{v}_4, \mathbf{v}_5, \mathbf{v}_6\}$ is a sublist of any such list, it is necessarily linearly independent.

Problem 4 **[MATALG]**

(a) Suppose that U is a matrix with orthonormal columns. Find the matrix V such that UV is the projection matrix onto the column space of U. Write your answer in the box.

(b) Suppose that $\mathbf{v}$ is orthogonal to every column of U. Find $U'\mathbf{v}$.

Solution

(a) The projection matrix onto the columns of U is UU', since U has orthonormal columns. Therefore, $V = U'$.

(b) $U'\mathbf{v}$ is equal to the zero vector, since each of its entries is a dot product of $\mathbf{v}$ with one of the columns of U.

Final answer:

U'

Problem 5 **[EIGEN]**

Find a 2×2 matrix A and a unit square S with sides not parallel to the axes such that A maps S to a long skinny rectangle with sides not parallel to the axes. Feel free to specify A using an unsimplified expression.

Solution

We define A to be a product of a rotation matrix, a diagonal matrix with one large diagonal entry and one small one, and another rotation matrix.

$$A = \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ \sqrt{2}/2 & -\sqrt{2}/2 \end{bmatrix} \begin{bmatrix} 100 & 0 \\ 0 & 0.01 \end{bmatrix} \begin{bmatrix} \sqrt{3}/2 & -1/2 \\ 1/2 & \sqrt{3}/2 \end{bmatrix}.$$

Problem 6 **[OPT]**

Which of the following functions f has the property that the gradient descent algorithm does not necessarily converge to the global minimum of f?

$$f(x) = x(1-x) \quad g(x) = -\frac{x^3}{3} + \frac{x^2}{2} \quad h(x) = x^2(1-x)^2$$

Solution

Gradient descent not converge for f because it isn't bounded below. Gradient descent cannot converge to the global minimum of g, since it has no minimum (because it's cubic). Gradient descent does not necessarily converge for h, because the derivative at $x = \frac{1}{2}$ is zero. However, if the point $x = \frac{1}{2}$ is never reached, the algorithm will converge to one of the two points where the global minimum is reached (0 and 1).

Problem 7 **[MATDIFF]**

Suppose that A is a matrix and $\mathbf{x}$ is a vector. Differentiate $|A\mathbf{x}|^2$ with respect to $\mathbf{x}$.

Solution

We have

$$|A\mathbf{x}|^2 = \mathbf{x}'A'A\mathbf{x},$$

so we apply the product rule and find that

$$\mathbf{x}'(A'A) + \mathbf{x}'(A'A) = \mathbf{x}'(A'A),$$

since the derivative of $B\mathbf{x}$ with respect to $\mathbf{x}$ is B (which we're applying here with $B = A'A$).

Final answer:

$$2\mathbf{x}'A'A$$

Problem 8 **[MACHARITH]**

Explain why `a == b` returns `false` if `a = 64 - 0.5^47 - 0.5^48` and `b = 64 - 0.5^48 - 0.5^47`

Solution

The tick spacing between 32 and 64 is $2^{-52} \cdot 2^5 = 2^{-47}$, so subtracting 2^{-47} from 64 results in the number one tick below 64. Subtracting a half tick spacing returns the value two tick spacings below 64, due to the round-to-even rule. Likewise, subtracting half a tick spacing from 64 yields 64 exactly, and then subtracting one tick spacing yields the number one tick spacing below 64.

Problem 9 **[NUMERROR]**

NumPy* has a function called `log1p` which returns the natural logarithm of the sum of 1 and the argument. Explain why such a function exists when you can achieve the same mathematical effect by just adding 1 and then taking the logarithm of the resulting sum. Hint: consider input values close to 0.

*R, Julia, MATLAB, Mathematica, and many other languages have such a function too.

Solution

The function $\log(1+x)$ is well-conditioned near $x = 0$ since its derivative is $1/(1+x) \approx 1$ for those values of x. However, the algorithm "add one then take the logarithm" is ill-conditioned because the condition number of $y \mapsto \log y$ is

$$\frac{y(1/y)}{\log y} = \frac{1}{\log y},$$

which is very large when near 1. Therefore, `log1p` is an implementation of some alternative algorithm which is stable.

Problem 10 **[PRNG]**

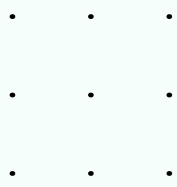
Consider a pseudorandom number generator which returns integer values in the interval $[0, 2^{32} - 1]$. Is it possible for the period of the PRNG to be greater than 2^{32}?

Solution

Yes, it is possible. If the PRNG is defined iteratively as $x_{n+1} = f(x_n)$ for some function f, then the period can be no greater than 2^{32}, since some number must appear for the second time in the first $2^{32} + 1$ terms. But an update rule defined in some other way can have a greater period.

Problem 11 **[COUNTING]**

How many triangles can be drawn with vertices at the dots shown?



Solution

There are $\binom{9}{3} = 84$ sets of three vertices, and $3 + 3 + 2 = 8$ of them are collinear and therefore do not define a triangle. Therefore, there are $84 - 8 = 76$ triangles.

Final answer:

76

Problem 12 **[PROBSPACE]**

Consider a random experiment which involves rolling a die, flipping a coin, and drawing a single card from a standard deck of 52 cards. Describe a probability space Ω for modeling this random experiment. How many elements does it have? Do the die and the coin correspond to disjoint subsets of Ω?

Solution

We take Ω to be the Cartesian product

$$\{1, 2, 3, 4, 5, 6\} \times \{\mathrm{H}, \mathrm{T}\} \times \{1, 2, \ldots, 52\}.$$

The number of elements in ω is $6 \cdot 2 \cdot 52 = 624$. The die and coin do **not** correspond to disjoint subsets of Ω. Every element of Ω has a component which corresponds to the die and a component which corresponds to the coin.

Problem 13 **[PMF]**

Suppose that a coin is weighted so that it is twice as likely to come up heads as tails. What is the probability mass function of the number of heads which appears in two independent flips of this coin?

Solution

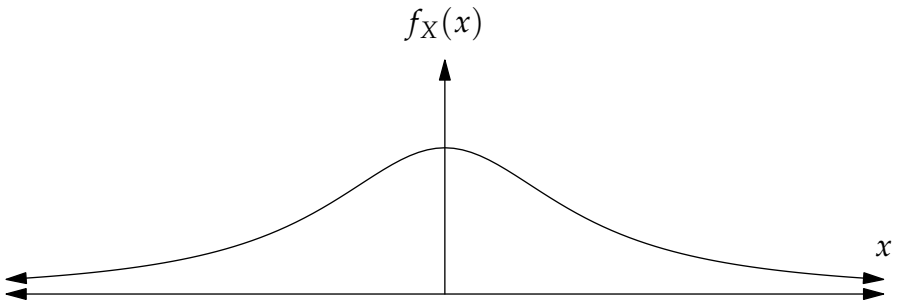
The probability that it comes up tails both times is $(1/3)^2 = 4/9$, and the probability it comes up heads both times is $(2/3)^2 = 4/9$. The probability that it comes up heads exactly once is therefore $1 - (1/3)^2 - (2/3)^2 = 4/9$. So the PMF of the number of heads assigns mass $1/9, 4/9$ and $4/9$ to the points 0, 1, and 2, respectively.

Problem 14 **[PDF]**

Consider a ray from the origin with angle (measured with respect to the positive x-axis) chosen uniformly between 0 and 180 degrees. Let X be the x-coordinate of the point where this ray intersects the line $y = 1$. Sketch the PDF of X.

Solution

Each short segment $[x, x + \Delta x]$ is more likely to contain X if it's closer to 0, since it corresponds to a longer interval of angles. Also, the PDF is clearly symmetric about 0. A graph is sketched below.



Problem 15 [CONDPROB]

Consider the following experiment: we roll a die, and if it shows 2 or less we select Urn A, and otherwise we select Urn B. Next, we draw a ball uniformly at random from the selected urn. Urn A contains one red and one blue ball, while urn B contains 3 blue balls and one red ball.

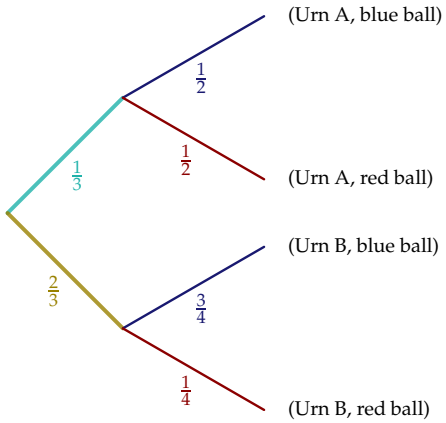
Find a probability space Ω which models this experiment, find a pair of events E and F such that $\mathbb{P}(E \mid F) = \frac{3}{4}$.

Solution

The four possible outcomes of this experiment are (A, blue), (A, red), (B, blue), and (B, red). So we let our probability space Ω consist of those four outcomes.

The probability of the outcome (A, blue) is equal to the probability that Urn A is selected times the conditional probability of selecting a blue ball given that Urn A was selected. We interpret the information that Urn A contains an equal number of blue and red balls as a statement that this conditional probability should be $\frac{1}{2}$. Therefore, we assign the probability $\frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$ to the event (A, blue).

Likewise, the probabilities we assign to the three other outcomes are $\frac{1}{6}$, $\frac{1}{2}$, and $\frac{1}{6}$, respectively.



With probabilities thus assigned to the outcomes in Ω, we should have $\mathbb{P}(E \mid F) = \frac{3}{4}$ where E is the event that we select a blue ball and F is the event that Urn B was selected. Let us check that this is indeed the case:

$$\frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)} = \frac{\frac{1}{2}}{\frac{2}{3}} = \frac{3}{4}.$$

Problem 16 [BAYES]

Suppose that $Z = X + Y$ where X is 0 with probability 99.99% and 1000 with probability 0.01%, Y is a standard normal random variable, and X and Y are independent. Which of the following is closest to the conditional probability that $X = 1000$ given that $Z \geq 997$.

$$0 \quad \frac{1}{4} \quad \frac{1}{2} \quad \frac{3}{4} \quad 1$$

Solution

Bayes theorem implies that X is extremely likely to be 1000 given that $X + Y \geq 997$:

$$\mathbb{P}(X = 1000) = \frac{(0.01\%)\mathbb{P}(Y \geq -3)}{(0.01\%)\mathbb{P}(Y \geq -3) + (99.99\%)\mathbb{P}(Y \geq 997)} \approx 1.$$

So the closest value is 1.

Problem 17 **[IND]**

Consider a fair die roll X. Are the events $\{X \leq 4\}$ and $\{X \in \{2, 4, 6\}\}$ independent?

Solution

The probability of the two events are $\frac{2}{3}$ and $\frac{1}{2}$, and the probability of their intersection is $\frac{1}{3}$. Since $\frac{2}{3} \cdot \frac{1}{2} = \frac{1}{3}$, yes, they are independent.

Problem 18 **[EXP]**

A standard deck of cards is the Cartesian product $\{2, 3, 4, 5, 6, 7, 8, 9, 10, \text{J}, \text{Q}, \text{K}, \text{A}\} \times \{\heartsuit, \diamondsuit, \clubsuit, \spadesuit\}$. The values $\heartsuit$, $\diamondsuit$, $\clubsuit$, and $\spadesuit$ are called *suits*.

A *royal flush* is a set of five cards which are the 10, J, Q, K, and A of the same suit. The probability of a randomly chosen set of 5 cards from the deck being a royal flush is $1/649740$.

Suppose that five poker hands are dealt from a well-shuffled deck. Is it possible that all five hands are royal flushes? Find the expected number of royal flushes among the five hands.

Solution

It is not possible that all five people have a royal flush, since there are only four suits. The expected number of royal flushes is $5/649740$, by linearity of expectation.

Problem 19 **[COV]**

Can two random variables X and Y have the property that $X - Y$ has a smaller variance than both X and Y? What has to be true about the sign of $\text{Cov}(X, Y)$ for this to be the case?

Solution

We determine that

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2\,\text{Cov}(X, Y).$$

Therefore, the variance $\text{Var}(X - Y)$ can only be smaller than the variance of X and smaller than the variance Y if $\text{Cov}(X, Y)$ is positive (and indeed, larger than half the larger of the two variances). This is possible; for example, consider the case where $X = Y$.

Problem 20 **[CONDEXP]**

Suppose that N is a geometric random variable with probability $\frac{1}{2}$, and suppose that $X_1, X_2, \ldots$ is a sequence of Poisson random variables with mean 6. Find the expected value of $X_1 + X_2 + \cdots + X_N$.

Solution

We apply the tower law:

$$\mathbb{E}[X_1 + X_2 + \cdots + X_N] = \mathbb{E}[\mathbb{E}[X_1 + X_2 + \cdots + X_N \mid N]] = \mathbb{E}[6N] = 12.$$

Problem 21 **[COMDISTD]**

Disprove the claim that the sum of two independent geometric random variables is a geometric random variable.

Solution

Consider two independent geometric random variables each with parameter $\frac{1}{2}$. The probability that their sum is 1 is zero, which is already incompatible with the claim that the sum is geometric.

Problem 22 **[COMDISTC]**

Which distribution has smaller variance: (a) the uniform distribution on $[0, 1]$ or (b) the distribution with density proportional to e^{-8x^2}?

Solution

The uniform distribution has variance $(b-a)^2/12 = 1/12$, while the given Gaussian distribution has variance σ^2 where $-8x^2 = -x^2/(2\sigma^2)$ for all $x \in \mathbb{R}$. In other words, the Gaussian has variance $\sigma^2 = 1/16$. So the distribution with the smaller variance is the Gaussian.

Problem 23 **[RVINEQ]**

Show that for any random variable X and for any positive u and t, we have $\mathbb{P}(X \geq u) \leq \mathrm{e}^{-tu}\mathbb{E}[\mathrm{e}^{tX}]$.

Solution

We have $\mathbf{1}_{x \geq u} \leq e^{t(x-u)}$ for all $x \in \mathbb{R}$, since the left-hand side is zero unless $x \geq u$, and in that case the right-hand side is at least 1. Taking the expectation of both sides yields the desired inequality.

Problem 24 **[CLT]**

If $X_1, X_2, \ldots$ is a sequence of independent random variables with common density $\frac{1}{\pi(1+x^2)}$ on $\mathbb{R}$, then the distribution of $\frac{1}{n}(X_1 + X_2 \cdots + X_n)$ has the same distribution as X_1. Are the hypotheses of the central limit theorem satisfied in this case?

Solution

No, because otherwise the distribution of the average would have to be approximately Gaussian. The issue is that this distribution has neither a mean nor a variance, since integrals defining the first and second moment of the given distribution diverge.

Problem 25 **[KDE]**

Discuss the behavior of the Nadaraya-Watson estimator when the bandwidth λ is very small and when λ is very large. Discuss the following phases: (1) λ is extremely small, (2) λ is not terribly small but smaller than the optimal value, (3) λ is not terribly large but larger than the optimal value, and (4) λ is extremely large.

Solution

(1) The Nadaraya-Watson estimator isn't even defined for many values of x when λ is extremely small. (2) As λ increases, the boxes of probability mass link up, but the estimator is still quite bumpy and too high-variance. When λ exceeds its optimal value, the regression function becomes too smooth, giving too much credence to points which are far away. As $\lambda \to \infty$, all points contribute equally, so the graph of the regression function converges to a horizontal line at the mean of the y-coordinates of the sample points.

Problem 26 **[LR]**

Explain how to find the function of the form

$$A + B \sin x + C \sin 2x + D \sin 3x$$

which has the least residual sum of squares for a given set of points $\{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$.

Solution

We form the matrix X which has a column of ones, a column with the values $\sin x_1, \sin x_2, \ldots, \sin x_n$, a column with the values $\sin 2x_1, \sin 2x_2, \ldots, \sin 2x_n$, and a column with the values $\sin 3x_1, \sin 3x_2, \ldots, \sin 3x_n$. Then we define $[A, B, C, D]$ to be $(X'X)^{-1}X'\mathbf{y}$, where $\mathbf{y} = [y_1, \ldots, y_n]$.

Problem 27 **[QDA]**

Consider applying QDA to a binary classification problem where it turns out that the plug-in estimators of the class conditional covariance matrices are exactly equal. What is the shape of the resulting decision boundary in that case?

Solution

QDA is equivalent to LDA in this case, which implies that the decision boundary is a hyperplane (a line, if the feature space is two-dimensional).

Problem 28 **[STATLEARN]**

Consider a supervised learning model which consists of a space $\mathcal{X} \times \mathcal{Y}$, a probability measure $\mathbb{P}$, and a loss functional L. Explain the differences between the loss functional for a prediction function h and the *empirical* loss (or empirical risk) of h.

Solution

The loss functional is a performance measure for the prediction function h and is evaluated with respect to the measure $\mathbb{P}$. Meanwhile, the empirical loss is an *estimate* of the loss functional obtained by evaluating the loss functional with respect to the empirical measure.

Problem 29 **[NPL]**

Consider a binary classification problem where the distribution of the positive class is equal to the distribution of $3 + X$ where X is an exponential random variable with parameter 1. Suppose further that the distribution of the negative class is equal to the distribution of an exponential random variable with parameter 1.

(a) What is the maximum possible detection rate?

(b) What is the lowest possible false alarm rate among those prediction functions whose detection rate is equal to your answer from (a)?

Solution

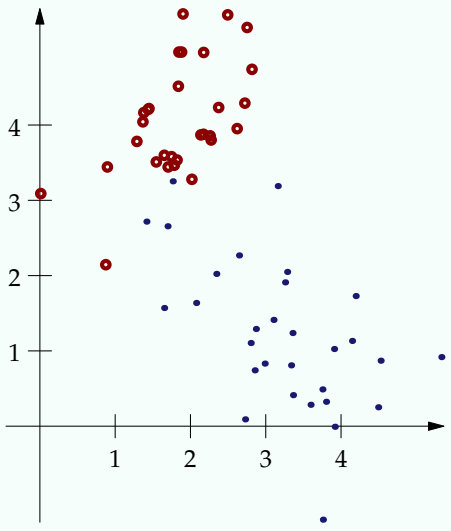
(a) The maximum possible detection rate is 100%, as always. We could just classify every input to $+1$.

(b) Since the positive class has no mass on $[0, 3]$, we can safely classify all of those inputs as negative. Therefore, the minimum possible false alarm rate is the amount of probability mass assigned to $[3, \infty)$ by $\text{Exp}(1)$, which is $\boxed{e^{-3}}$.

Problem 30 **[SVM]**

(a) Find the minimum number of points that would have to be discarded to apply hard-margin SVM (directly; no kernel function) to the classification problem shown.

(b) Find a choice of decision boundary and margin width which minimizes the *number* of points with a nonzero contribution to the soft-margin SVM loss estimator. Sketch your slab and indicate the set of points with nonzero loss contribution.



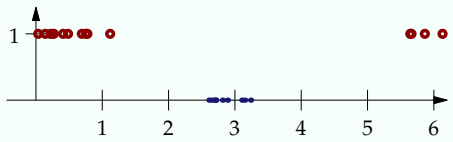
Solution

(a) We can drop two points: the red one near $(1, 2)$ and the blue one near $(2, 3)$ (let's call them A and B). It is not possible to do it with one point, since it's clear that leaving either A or B would require removing more points of the opposite color.

(b) This is kind of the same question: we can draw a very thin slab which separates the red points from the blue points except for A and B, and in that case only those two points make a positive contribution to the loss estimator.

Problem 31 [LOGIST]

Consider the one-dimensional classification problem shown. The feature space is $\mathbb{R}^1$, indicated by the x-axis, and the response variable is displayed along the y-axis.



(a) Explain why logistic regression will not be able to accurately model the conditional probability of class 1 in this scenario.

(b) Sketch your best guess of the graph of the function which minimizes the logistic regression loss function. Explain your considerations for drawing the graph the way you did.

Solution

(a) Applying σ^{-1} to the regression function $r(\mathbf{x})$ does not yield a function which is approximately linear. This stems from the red points being in two clusters which are on either side of the blue cluster.

(b) The best fit logistic regression function will decrease gradually over the interval shown in the figure, because if it had a steep slope, then it would pay a very large penalty for badly misclassifying the second cluster of red points.

Problem 32 [DR]

Consider a set of 10,000 vectors in $\mathbb{R}^{784}$, each representing a handwritten digit image. Let $\mathbf{v}_k$ be the kth principal component of this set of vectors. Describe the differences between the point cloud obtained by applying $\mathbf{x} \mapsto (\mathbf{v}_1 \cdot \mathbf{x}, \mathbf{v}_{100} \cdot \mathbf{x})$ to all of the vectors and the one obtained by applying $\mathbf{x} \mapsto (\mathbf{v}_{200} \cdot \mathbf{x}, \mathbf{v}_{201} \cdot \mathbf{x})$ to them.

Solution

The first point cloud will be very long in the horizontal direction and slender in the vertical direction. The second point cloud will be quite round, with both dimensions smaller than the shorter dimension of the first point cloud.

Problem 33 **[NN]**

Consider a neural network with ReLU activations and two affine maps: $A_1(\mathbf{x}) = \begin{bmatrix} 2 & -3 & 4 \\ 0 & -4 & 0 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 3 \\ 0 \end{bmatrix}$ and $A_2(\mathbf{x}) = \begin{bmatrix} -6 & 1 \\ 4 & 5 \\ 0 & 1 \\ 2 & -2 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 2 \\ -1 \\ 5 \\ 3 \end{bmatrix}$.

Consider a sample with $\mathbf{x}_i = \begin{bmatrix} 3 \\ 1 \\ -1 \end{bmatrix}$ and $y_i = \begin{bmatrix} 1 \\ -2 \\ -1 \\ 0 \end{bmatrix}$. Find the suggested change to the weight matrix W_1 assuming a learning rate of $\epsilon = 0.1$.

Solution

The forward propagated vectors are $\begin{bmatrix} 2 \\ -4 \end{bmatrix}$, $\begin{bmatrix} 2 \\ 0 \end{bmatrix}$, and $\begin{bmatrix} -10 \\ 7 \\ 5 \\ 7 \end{bmatrix}$. Then backwards propagating yields (right to left) $[-22\ 18\ 12\ 14]$, $[232\ 52]$, and $[232\ 0]$. Finally, the derivative of the cost with respect to W_1 is

$$\begin{bmatrix} 232 \\ 0 \end{bmatrix} \begin{bmatrix} 3 & 1 & -1 \end{bmatrix},$$

so the suggested change is $\begin{bmatrix} -69.6 & -23.2 & 23.2 \\ 0 & 0 & 0 \end{bmatrix}$.

Problem 34 [R]

Write an R function that maps a 2×2 rotation matrix to the angle of rotation in degrees. For simplicity, you may assume that the angle is strictly between 0 and 90 degrees. The inverse tangent function in R is called `atan`, and it returns values in radians.

```
near(angle(matrix(c(cos(pi/4),sin(pi/4),-cos(pi/4),sin(pi/4)),2,2)), 45)
```

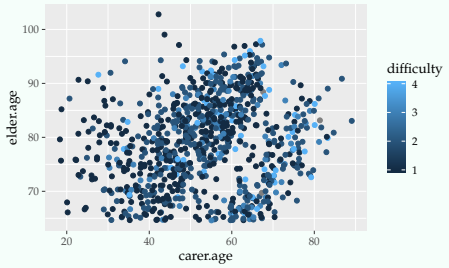
Solution

R

```
angle <- function(m){180/pi*atan(m[2,1]/m[1,1])}
```

Problem 35 [GGPLOT]

Write `ggplot2` code to generate the plot shown for the data frame `elder.care`. Hint: the geom isn't actually `geom_point`, but rather a closely related one. Also, set the size of the points to 1.



Solution

It's important that the size argument appear outside `aes`, since it's a setting and not a mapping of data to aesthetics.

```
ggplot() +
  geom_jitter(aes(carer.age,elder.age,color=difficulty),size=1)
```

Problem 36 [DPLYR]

Consider a data frame storing the standards-based medal information for a class, with column names "student.name", "type", and "standard", "medal", with an example record of "Jane Doe", "Overall", "LOGIST", "Silver", or "John Doe", "Homework", "SVM", "Bronze". Write `dplyr` code to return a data frame such that each row specifies a standard and the number of students who have an overall gold for that standard and the number of students who have an overall silver for that standard. Call the dataframe `grades`.

Solution

```
grades %>%
  filter(type == "Overall") %>%
  group_by(standard) %>%
  summarise(golds = sum(medal == "Gold"), silvers = sum(medal == "Silver"))
```

Problem 37 **[POINTEST]**

Suppose that T is a statistical functional, ν is a probability measure on $\mathbb{R}$, $\widehat{\theta}_1$ is a biased estimator of $\theta = T(\nu)$, and $\widehat{\theta}_2$ is a unbiased estimator of $\theta = T(\nu)$. Is it possible that the MSE of $\widehat{\theta}_1$ is smaller than the MSE of $\widehat{\theta}_2$, even though the former is biased and the latter is not? Explain why it is not possible or give an example (you do not have to do any calculations to support your example; you can just cite it).

Solution

Yes, it is possible that a biased estimator has smaller MSE. We saw an example on the homework: if we estimate the maximum of distribution from the family $\{\text{Unif}([0,\theta]) : \theta > 0\}$ with the plug-in estimator and with twice the sample mean, then the former is negatively biased but has smaller MSE.

Problem 38 **[BOOT]**

Consider the statistical functional $T(\nu)$ which returns the expectation of the minimum of two independent samples from ν.

(a) Find the *exact limiting value* of the bootstrap estimate of $T(\nu)$ given the samples

$$3, 4, 3, 7$$

from ν.

(b) Even though you found the exact value of the bootstrap estimate of $T(\nu)$, is that value necessarily close to $T(\nu)$?

Solution

(a) By the law of large numbers, the bootstrap estimate converges to the expectation of $\min(A, B)$, where A and B are independent draws from the list $[3, 4, 3, 7]$. The 16 elements of the Cartesian product $[3, 4, 3, 7]^2$ are mapped by the random variable $\min(A, B)$ to the 16 values $[3, 3, 3, 3, 3, 4, 3, 4, 3, 3, 3, 3, 3, 4, 3, 7]$. Therefore, the limiting value is $55/16$ (the average of these values).

(b) No, $T(\nu)$ is not necessarily close to $T(\widehat{\nu})$. Typically it would be close if many samples from ν are available, but in this case we only have four.

Problem 39 **[HYPTEST]**

A *zero-knowledge* proof is a method by which one person can demonstrate knowledge to another person without revealing anything beyond the fact of that knowledge.

For example, consider a red ball, a green ball, and a color blind friend who is skeptical that the balls are actually distinguishable. It is possible to convince your friend that you can distinguish the balls without revealing which ball is red and which is green: have the friend hold both balls behind his back and choose one of them randomly to reveal to you. Then he puts both balls behind his back again and reveals a ball for a second time, switching them behind his back with probability $1/2$. You then indicate whether he switched the balls.

Describe this random experiment in a hypothesis test framework. Identify the null hypothesis and the alternative hypothesis. Determine the number of times the experiment must be repeated to reject the null hypothesis with 99% confidence (bearing in mind that the balls are actually different colors, so you will in fact be able to distinguish them every time).

Solution

The null hypothesis is that the balls are indistinguishable, and the alternative hypothesis is that they are distinguishable. The probability of successfully determining whether the balls are switched k times is $1/2^k$. So to reject the null hypothesis at a 1% significance, we need $1/2^k < 0.01$. The least value of k for which this inequality holds is $k = 7$.

Problem 40 **[MLE]**

Consider a Bernoulli distribution with unknown parameter p. Show that the maximum likelihood estimator of p is equal to the proportion of 1's.

Solution

The likelihood associated with $X_1, \ldots, X_n$ is

$$p^k(1-p)^{n-k},$$

where k is the number of ones in the list $\{X_1, \ldots, X_n\}$. So the log likelihood is

$$k \log p + (n-k)\log(1-p),$$

and differentiating with respect to p yields

$$\frac{k}{p} - \frac{n-k}{1-p},$$

which equals zero when $p = k/n$. Thus the MLE of p is equal to the proportion of ones.