Brown University
DATA 1010
Fall 2018: Midterm III
Samuel S. Watson

Name:

You will have three hours to complete the exam, which consists of 36 questions. Among the first 24 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]

Suppose that g and h are functions such that the range of h is a subset of the domain of g and that A is a subset of the domain of h with the property that $(g \circ h)(A) = \emptyset$.

(a) If f is a function whose domain is the range of g, then what is $(f \circ g \circ h)(A)$? Write your answer in the box.

(b) Explain why it needed to be stipulated that the range of h is a subset of the domain of g.

Solution

(a) We have $(f \circ g \circ h)(A) = \emptyset$, since there are no elements of the codomain of g with a preimage in the domain of h if there aren't any elements in the codomain of g with a preimage in the domain of h. (In fact, $A = \emptyset$).

(b) If the codomain of h is not a subset of the domain of g, then it is not necessarily possible to compose the functions, since values might be output by h which cannot be input into g.

Final answer:

$\emptyset$

Problem 2 [JULIA]

Write a Julia function which accepts a positive integer argument n and returns a vector of the next n leap years (starting from the current year, 2018). Bear in mind that a year is a leap year if it is a multiple of 400 **or** if it is a multiple of 4 which is not also a multiple of 100. So, for example, 2000 was a leap year, but 2100 will not be.

Solution

We take the approach of incrementing the year and appending it to the list of years until we've found n of them.

This could be made faster, for example by finding the next multiple of 4 and then incrementing by 4 rather than by 1. However, this approach is simple, and it gets the job done.

```julia
function leapyears(n,startingyear=2018)
    year = startingyear
    years = []
    while length(years) < n
        if year % 400 == 0 || (year % 4 == 0 && year % 100 ≠ 0)
            push!(years, year)
        end
        year += 1
    end
    years
end
```

Problem 3 **[LINALG]**

Give an example of a basis of $\mathbb{R}^5$ with the property that the first three vectors are pairwise orthogonal, but the last two vectors are orthogonal neither to one another nor to any of the first three vectors.

Solution

There are many such examples, but one simple approach is to take the first three vectors to be the first three standard basis vectors, and then define $\mathbf{v}_4 = [1, 1, 1, 1, 0]$ (which has a dot product of 1 with each of the first three vectors) and $\mathbf{v}_5 = [1, 1, 1, 1, 1]$ (which has a dot product of 4 with $\mathbf{v}_4$ and 1 with each of the first three).

Problem 4 **[MATALG]**

(a) Suppose that U is a 3×2 matrix with orthonormal columns. Suppose that $\mathbf{b}$ is in the range of U. Find the dimension of the solution set of the equation $UU'\mathbf{x} = \mathbf{b}$.

(b) Describe the geometric relationship between the solution set of $UU'\mathbf{x} = \mathbf{b}$ and the range of U.

Solution

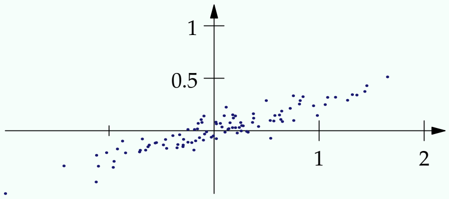
(a) The nullity of UU' is 1 by the rank-nullity theorem since its rank is equal to 2. (Recall that the rank of A is equal to the rank of AA' for any matrix A).

(b) Since UU' is the projection matrix onto the column space (that is, the range) of U, the set of points project to a given vector is **orthogonal** to the column space of U.

Problem 5 [EIGEN]

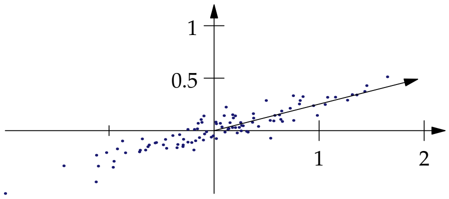
(a) Sketch the line through the origin with the property that its sum of squared distances to the points shown is as small as possible.

(b) Suppose you received a matrix containing the coordinates of the points shown as its columns. Explain what you would do (computationally) with the matrix to find the line requested in (a). You may assume you have standard linear algebra functions at your disposal.



Solution

The line is the span of the vector shown in the figure. We would determine this line from a matrix with the points as columns as the span of the first column of V in its singular value decomposition.



Problem 6 [OPT]

The solution of the ordinary least squares regression problem may be obtained using the explicit formula $(X'X)^{-1}X'\mathbf{y}$, where X is the matrix of regressors and $\mathbf{y}$ is the vector of training responses.

However, it is also possible to use a standard optimization approach. Consider the task of minimizing the squared length of $\mathbf{y} - X\boldsymbol{\beta}$. Explain how to apply gradient descent to this problem.

Solution

We begin at some arbitrary initial value for $\boldsymbol{\beta}$, differentiate $|\mathbf{y} - X\boldsymbol{\beta}|^2$ with respect to $\boldsymbol{\beta}$, and increment $\boldsymbol{\beta}$ by $-\epsilon$ times this derivative value. We repeat until the gradient becomes falls below a small, predetermined threshold.

Problem 7 **[MATDIFF]**

Consider the function $f : \mathbb{R}^{2\times 2} \to \mathbb{R}$ defined by $f\left(\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]\right) = a + b + c + d$. Differentiate $f(A)$ with respect to A.

Solution

We differentiate the expression for $f(A)$ with respect to a (which yields 1), and place the result in the $(1, 1)$ position. Continuing in this way, we find the derivative is a 2×2 matrix of ones.

Final answer:

$$\left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right]$$

Problem 8 **[MACHARITH]**

Consider the `Float64` number system with the subnormal numbers removed. Give an example to show that the implication `x - y == 0` $\Longrightarrow$ `x == y` is false in that system.

Solution

If we subtract the smallest two distinct normal numbers (2^{-1022} and $2^{-1022} + 2^{-1074}$), the difference between them is much closer to 0 than to any normal numbers and therefore must be rounded to zero. Thus these numbers are unequal, yet differencing them returns zero in this number system.

Problem 9 **[NUMERROR]**

Your friend says to you "that function is unstable near $x = 2$, so we should use a well conditioned algorithm if we want to evaluate it accurately for those values of x". Explain your friend's mistakes.

Solution

First, a function is well-conditioned or ill-conditioned, not stable or unstable. Likewise, an algorithm is stable or unstable, not well-conditioned or ill-conditioned. So those words should be switched.

Next, using a stable algorithm does not ensure that we will evaluate the function's values accurately. It merely ensures accuracy *relative to the condition number*. If the function is not well-conditioned, using a stable algorithm will yield high levels of accuracy. It is the opposite situation, when the function is well conditioned but the algorithm is unstable, where we can rectify the situation by using a stable algorithm.

Problem 10 **[PRNG]**

(a) Given the output of a pseudorandom number generator (which outputs integers between 0 and $2^{64} - 1$), we split the terms of the sequence into blocks of 3 and check for each block whether its terms are in increasing order. We do this for the first 3000 terms and find that approximately 16.5% of the blocks are in order. Did this test distinguish the sequence from a genuinely random sequence?

(b) Does this test ensure that the sequence does behave sufficiently like a random sequence for use in scientific computing applications?

Solution

Since there are $3! = 6$ possible orderings, we would expect about $\frac{1}{6}$ of the triples to be in increasing order. So this test does not distinguish from a genuinely random sequence. On the other hand, many other statistical tests would have to be applied to conclude that the PRNG is suitable for use in applications. Just one test is hardly sufficient.

Problem 11 **[COUNTING]**

You have 5 colors of stationery, 4 types of envelopes, and 2 stamps. The blue stationery doesn't go with the purple envelopes, but otherwise you can use whatever combinations you want. How many ways are there to combine a stationery color with an envelope and a stamp?

Solution

There are $5 \times 4 \times 2 = 40$ total combinations, and of them $1 \times 1 \times 2 = 2$ combinations are excluded. So there are $40 - 2 = \boxed{38}$ permitted combinations.

Final answer:

38

Problem 12 **[PROBSPACE]**

Explain why it is not true in general that $P(A \setminus B) = P(A) - P(B)$. Provide conditions for which it is true, and explain why it is true under your stated conditions.

Solution

One way to see that it is not true in general is that the right-hand side can be negative: choose any events A and B such that B is more likely than A. Since $P(A \setminus B) \geq 0$, the given equation is necessarily false.

The equation is true if $B \subset A$. In that case, it follows from additivity: $P(B) + \mathbb{P}(A \setminus B) = \mathbb{P}(A) \implies P(A \setminus B) = P(A) - P(B)$.

Problem 13 [PMF]

Suppose that Ω has three elements with masses $\frac{1}{6}$, $\frac{2}{6}$, and $\frac{3}{6}$. Describe the set of possible distributions for a random variable $X : \Omega \to \mathbb{R}$.

Problem 14 [PDF]

Find the PDF of U^2, where U is a random variable distributed uniformly on $[0, 1]$.

Solution

We have $P(U^2 < t) = \mathbb{P}(U < \sqrt{t}) = \sqrt{t}$ for all t between 0 and 1, so the PDF of U^2, which is the derivative of its CDF, is the derivative of $\sqrt{t}$, which is $\frac{1}{2\sqrt{t}}\mathbf{1}_{0 \leq t \leq 1}$.

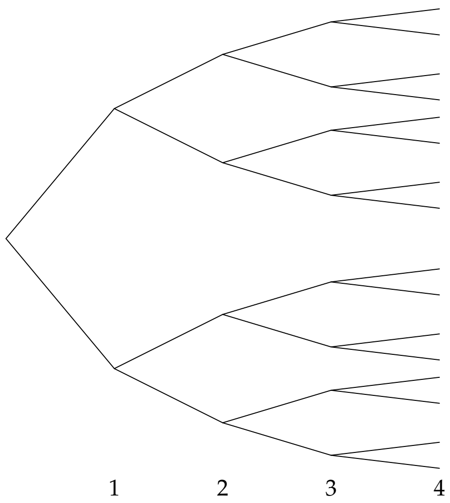
Problem 15 **[CONDPROB]**

Make a branching tree diagram to illustrate a sequence of four fair, independent coin flips. Show, in the context of the diagram, that the conditional probability of getting heads on the third flip given that the second flip was tails, is $\frac{1}{2}$.

Solution

The tree diagram is as shown, with upwards arrows indicating heads and downwards arrows indicating tails. The set of endpoints on the far right of the diagram is Ω.

The second flip is tails for the fourth through eighth and the thirteenth through sixteenth endpoints (counting from the top). Of these, the ones corresponding to heads on the third flip are the fourth, fifth, thirteenth, and fourteenth. Therefore, the conditional probability of getting heads on the third flip given that the second flip is tails is $\frac{4/16}{8/16} = \frac{1}{2}$.



Problem 16 **[BAYES]**

A battery manufacturer has three factories which produce 35%, 35%, and 30% of the batteries produced by the company. The three factories have defectiveness rates of 1.5%, 1%, and 2%, respectively.

Given that a randomly selected battery is defective, write a numerical expression for the conditional probability that the defective battery came from the second factory. You do not have to evaluate the expression.

Solution

The probability that the battery is defective is

$$[0.35, 0.35, 0.3] \cdot [0.015, 0.01, 0.02].$$

The portion of this probability mass which corresponds to the second factory is $(0.35)(0.01)$. Therefore, the conditional probability that the battery is from the second factory, given that it is defective, is

$$\boxed{\frac{(0.35)(0.01)}{(0.35)(0.015) + (0.35)(0.01) + (0.3)(0.02)}} = 23.7\%$$

Problem 17 **[IND]**

Give an example of a discrete probability space Ω satisfying the following conditions:

(i) No two elements of Ω have the same probability mass,

(ii) It is possible to define two independent random variables on Ω, each of which has Bernoulli distribution with parameter $p = \frac{1}{2}$.

What is the smallest number of elements that such a probability space could have? Write your answer in the box.

Solution

The joint distribution of X and Y is the uniform measure on the four-element set $\{0,1\}^2$. Therefore, the masses of Ω must get mapped to these four elements in such a way that the total mass mapped to each point is $\frac{1}{4}$. Since the masses in Ω are required to be distinct, we must split at least three of the one-quarter masses in different ways.

For example, we could let $\Omega = \{1,2,3,4,5,6,7\}$, with mass $\frac{1}{8}$ at 1 and at 2, with mass $\frac{1}{16}$ and $\frac{3}{16}$ at 3 and 4 respectively, with mass $\frac{1}{32}$ and $\frac{7}{32}$ at 5 and 6 respectively, and with mass $\frac{1}{4}$ at 7.

Final answer:

7

Problem 18 **[EXP]**

Give an example showing that is possible to calculate the expectation of a random variable *without* working out the distribution of the random variable.

Solution

Such a feat may be achieved using linearity of expectation. For example, we can find the expected number of players who select the correct jersey, if n players on a team put on their n jerseys at uniformly at random: the probability that the first player gets their jersey is $\frac{1}{n}$, and similarity for the other players. Therefore, if we let X_i be the indicator of the event that the ith player gets their jersey, then

$$\mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = n(1/n) = 1$$

Problem 19 **[COV]**

(a) True or false: random variables whose joint distribution is multivariate normal have covariance zero

(b) True or false: given that X and Y each have unit variance, the set of possible values of $\operatorname{Cov}(X, Y)$ is the interval $[-1, 1]$.

(c) True or false: if the joint distribution of (X, Y) has positive mass only at the points $(0, 1)$, $(1, 0)$ and $(-1, -1)$, then X and Y necessarily have positive covariance.

Solution

(a) False. If the off-diagonal entries of Σ are nonzero, the components of a multivariate normal random vector are correlated.

(b) True. The correlation is always between -1 and 1, and correlation is equal to covariance if the random variables have a product of standard deviations equal to 1.

(c) False. If enough of the mass is at $(1, 0)$ and $(0, 1)$ (and sufficiently equally distributed), then they will be negatively correlated.

Problem 20 **[CONDEXP]**

Suppose that X and Y have joint PDF given by $f(x, y) = x + y$ on the unit square $[0, 1]^2$. Find the conditional expectation of Y given X.

Solution

We have

$$\mathbb{E}[Y \mid X] = \frac{\int_0^1 y(x+y)\,\mathrm{d}y}{\int_0^1 (x+y)\,\mathrm{d}y} = \frac{\frac{x}{2} + \frac{1}{3}}{x + \frac{1}{2}} = \frac{3x+2}{6x+3}.$$

Final answer:

$\frac{3x+2}{6x+3}$

Problem 21 [COMDISTD]

(a) Suppose that X is Poisson distributed and Y is a Bernoulli random variable. Is it possible that $X = Y$? Is it possible that X and Y are independent? Answer these two questions by writing YY, YN, NY, or NN in the box.

(b) The distribution of the sum of a Poisson(λ_1) random variable and an independent Poisson(λ_2) random variable is a named distribution. Identify it. You do not have to prove that this is the case, but you will be able to reason pretty confidently by considering the distribution that the Poisson distribution approximates.

Solution

(a) **NY**. Random variables cannot be equal unless they have the same distribution. However, any pair of distributions is possible for independent random variables.

(b) The sum is Poisson$(\lambda_1 + \lambda_2)$, because Poisson(λ_1) is approximately binomially distributed with parameters n and λ_1/n for some large n, and similarly for Poisson(λ_2). By adding up the $2n$ random variables in pairs, we can think of this sum of binomial random variables as a sum of n random variables which are equal to 1 with probability $(\lambda_1 + \lambda_2)/n - \lambda_1\lambda_2/n^2$, equal to 2 with probability $\lambda_1\lambda_2/n^2$, and equal to 0 with the remaining probability. For very large n, these random variables are approximately Bernoulli with success probability $(\lambda_1 + \lambda_2)/n$. Applying the Poisson approximation again, we conclude that the sum should be Poisson$(\lambda_1 + \lambda_2)$ (and indeed, it is).

Final answer:

NY

Problem 22 [COMDISTC]

Show that the exponential distribution is *memoryless*, meaning that $\mathbb{P}(X > t + s \mid X > t) = \mathbb{P}(X > s)$ if X is exponentially distributed and if s and t are nonnegative real numbers.

Solution

We have

$$\mathbb{P}(X > t + s \mid X > t) = \frac{\mathrm{e}^{-\lambda(s+t)}}{\mathrm{e}^{-\lambda t}} = \mathrm{e}^{-\lambda s} = \mathbb{P}(X > s),$$

as desired.

Problem 23 **[RVINEQ]**

Markov's inequality states that if X is a random variable taking values in $[0, \infty)$, then $\mathbb{P}(X \geq a) \leq \mathbb{E}[X]/a$, for all $a > 0$. Prove Markov's inequality by showing that $\mathbf{1}_{\{x \geq a\}} \leq x/a$ for all $x \geq 0$ and then substituting $X(\omega)$ for x and taking the expectation of both sides.

Solution

We have $\mathbf{1}_{\{x \geq a\}} \leq x/a$ for $0 \leq x < a$ since the left-hand side is zero in that case, and it's true when $x \geq a$ since the right hand side is at least 1 in that case. Substituting X for x and taking the expectation, we get $\mathbb{P}(X \geq a) \leq \mathbb{E}[X/a] = \mathbb{E}[X]/a$, as desired.

Problem 24 **[CLT]**

A fair coin is flipped 10,000 times. Approximate the probability of the event that the coin turns up heads more than 5100 times.

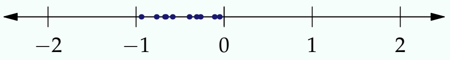
Solution

By the central limit theorem, the distribution is approximately normal with mean 5000 and standard deviation $\sigma\sqrt{n} = \frac{1}{2} \cdot 100 = 50$. Therefore, the desired probability is approximately the amount of probability mass lying two standard deviations above the mean for a normal distribution, which is about 2.3%.

Final answer:

2.3%

Problem 25 [KDE]



(a) Which of the following is a reasonable choice of bandwidth λ for a one-dimensional KDE with the samples shown?

(a) $\lambda = -1$

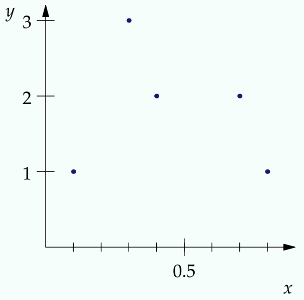
(b) $\lambda = 0.25$

(c) $\lambda = 2$

(d) $\lambda = 10$

(e) $\lambda = 20$

(b) For what values of λ is the Nadaraya-Watson estimator r_λ defined at $x = \frac{1}{2}$? (Note: every point in the figure is contained a vertical line through one of the tick marks.)



Solution

(a) $\lambda = 0.25$ is the most reasonable choice. Negative values don't make sense, and the other three values would all result in a very flat distribution with a lot of mass far away from the interval where all of the points lie.

(b) The estimator is defined when the corresponding density estimator places a positive amount of mass along the line $x = 0.5$, which is true for all values $\lambda > 0.1$, since the nearest point from $x = 0.5$ is 0.1 units away.

Problem 26 [LR]

(a) Find the minimum residual sum of squares for a linear prediction function for the samples

$$((x_1, x_2), y) = ((11, 7), 19), ((1000, 26), 1027), ((14, -2), 13), ((7, 100), 108), ((-4, -5), -8), ((0, 1), 2), ((6, 0), 7)$$

(b) Suppose that ordinary least squares regression is applied with quadratic combinations of x_1 and x_2. What will the coefficients of the quadratic terms be?

Solution

(a) The best prediction function is $y = x_1 + x_2 + 1$, which has zero RSS.

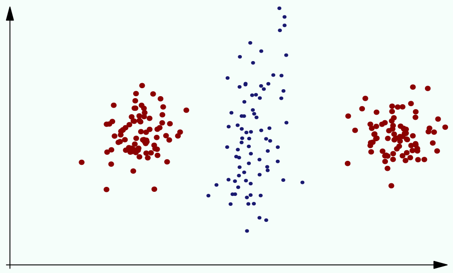
(b) The quadratic coefficients are zero, since the linear terms are sufficient to fit the data perfectly (and there are enough data points to uniquely specify a unique degree-2-or-lower polynomial which fits the data points).

Problem 27 [QDA]

(a) Make a case for **not** using QDA for the classification problem shown.

(b) Is it possible that QDA does nevertheless manage a 100% training accuracy?

Hint: draw your own decision boundary and ask yourself whether that sort of boundary is a possible QDA boundary.



Solution

(a) QDA is based on the assumption that the class conditional distributions are multivariate Gaussian, and the red one is bimodal and therefore clearly not.

(b) Yes, QDA can achieve a hyperbolic decision boundary which does separate the red and blue points.

Problem 28 [STATLEARN]

(a) Your colleague proposes the following machine learning algorithm, which is based on 20 pre-selected machine learning algorithms (all applicable to the same problem type).

Given a problem of the appropriate type, we split the available data into a training set and a test set. For each of the 20 algorithms, we train them on the training set and test the resulting fit model on the test set. We then return the prediction of the algorithm with the least test error.

How would you expect the resulting predictor to perform on unseen data?

(b) Consider a supervised learning problem with feature space $\mathcal{X} = \{1, 2\}$ and response space $\mathcal{Y} = \{+1, -1\}$. What probability measure on $\mathcal{X} \times \mathcal{Y}$ maximizes the misclassification probability for the Bayes classifier (in other words, what is the worst-case scenario for classification accuracy)?

Solution

(a) The problem with this algorithm is that it uses so-called test data in a training role. Therefore, one should expect the model to be overfit.

Said another way, some models might not have especially low generalization error but might have unusually low error on this particular test set, purely by chance. If we test many models, the chances are high that this will occur for at least one of them, and then we will be stuck using that one instead of one of the others with better generalization error.

(b) One such measure is the uniform measure. Then each classification will be a toss-up, and the Bayes classifier will have 50% misclassification probability.

Problem 29 [NPL]

Consider a binary classification problem with $f_+(x) = \mathbf{1}_{\{0 \le x \le 1\}}$ and $f_-(x) = \mathbf{1}_{\{\frac{2}{3} \le x \le \frac{5}{3}\}}$, and $p_+ = \frac{1}{4}$ and $p_- = \frac{3}{4}$. Find all points of the form $(\mathrm{FAR}(h_t), \mathrm{DR}(h_t))$, where $t \in (0, \infty)$ and where h_t is the likelihood ratio classifier. Hint: there are only two!

Write your answer in the box in the form $\{(a, b), (c, d)\}$, where $a, b, c, d \in \mathbb{R}$.

Solution

As x ranges over the interval $[0, 5/3]$, there are only three possible values for the ratio $f_+(x)/f_-(x)$, namely 0, 1, and ∞. So if t is between 0 and 1, we classify as $+1$ for values of x between 0 and 1 and -1 otherwise, and if t is greater than 1, we classify as $+1$ when x is between 0 and $\frac{2}{3}$ (and -1 otherwise).

In the former case, the false alarm rate is $\frac{1}{3}$, since two-thirds of the mass of the negative distribution lies in a region which classifies as $+1$. The detection rate is 1, since we always classify as $+1$ in the region where the positive class conditional distribution has nonzero density.

In the latter case, the false alarm rate is 0, since the whole negative distribution lies in a region classified as -1. The detection rate is $\frac{2}{3}$, since that's how much of the positive-distribution mass that lies in the positive-classification region.

Final answer:

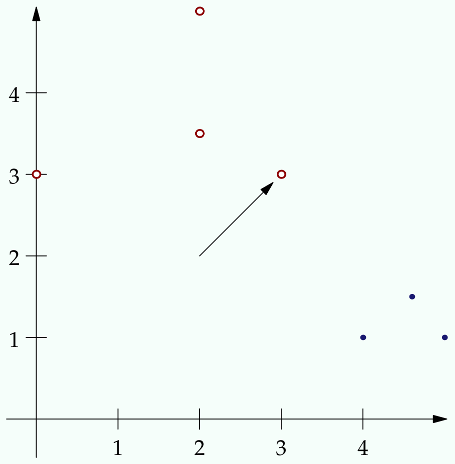
$$\{(\tfrac{1}{3}, 1), (0, \tfrac{2}{3})\}$$

Problem 30 [SVM]

(a) How does it change the slope of the decision boundary if the indicated point is dropped from the data set? Sketch your answer.

(b) Consider a soft-margin SVM classifier with $|\boldsymbol{\beta}| = 1$ and decision boundary given by the horizontal line with intercept $(0, 2.5)$. How many of the 7 points have a **positive** contribution to the SVM loss estimator, given that not all of the points have a positive contribution? (Note: the vertical coordinate of every point in the figure is an integer multiple of $\frac{1}{2}$.) Write your answer in the box.

(c) When λ is sufficiently large, the SVM loss minimizer will predict the same classification for all 7 points. Which class is it (red hollow or blue solid)?

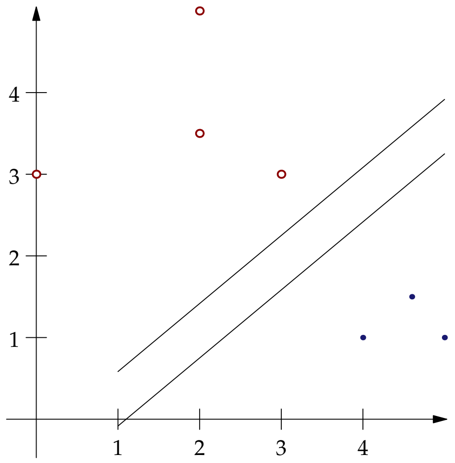


Solution

(a) The slope of the decision boundary does not change. The separating slab increases in thickness, and the decision boundary moves up. The two decision boundaries are shown in the figure to the right.

(b) The **two** hollow red points with an ordinate of 3 have a positive contribution to the loss function. All other points are on the edge of the slab or beyond, on the correct side.

(c) When λ is large, the optimization is forced to use a large slab thickness, so it will tend to move away from the class which has more points. Therefore, in this case it will classify them all of the points as red hollow.



Final answer:

2

Problem 31 [LOGIST]

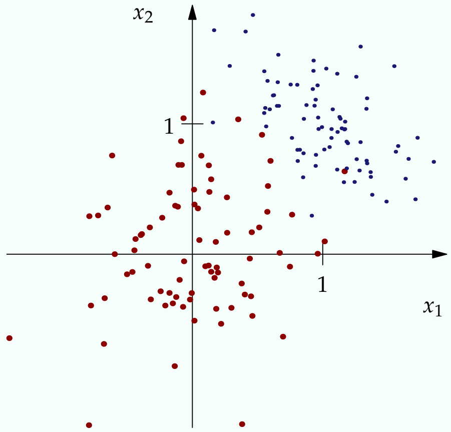
(a) Which of the following prediction functions has the least logistic regression loss estimator value? (In other words, which is the best logistic regression fit?) Assume that the small blue dots are class $+1$ and the larger red dots are class -1.

$$\sigma(x_1 - 2) \quad \sigma(x_1 + 1) \quad \sigma(-x_1 + x_2 + 1)$$

$$\sigma(x_1 - x_2 + 1) \quad \sigma(-x_1 - x_2 + 1) \quad \sigma(x_1 + x_2 - 1.5)$$

Here σ denotes the logistic function $x \mapsto 1/(1 + e^{-x})$.

(b) Consider scaling all of the coefficients inside the parentheses in your answer to (a) by a factor of 10^6 (including the constant term). Would the decision boundary be the same? Would the resulting fit be better or worse (from the point of view of the logistic regression loss estimator)?



Solution

(a) The one which fits the data the best is $\sigma(x_1 + x_2 - 1.5)$, since the direction orthogonal to $x_1 + x_2 - 1.5 = 0$ (which is $[1, 1]$, the vector of coefficients of x_1 and x_2) runs in the direction along which the two classes are maximally distinguished. Also, its decision boundary $x_1 + x_2 - 1.5 = 0$ does pass through the overlap of the two point clouds.

(b) Scaling all of the coefficients will preserve the boundary, but it will make all of the predictions much more confident. Since the logistic regression loss includes a penalty for overconfident, wrong predictions, this will increase the loss estimator and make the fit worse.

Final answer:

$\sigma(x_1 + x_2 - 1.5)$,
worse

Problem 32 [NN]

(a) Consider a neural network with biases all zero. Suppose that a given input vector $\mathbf{x}$ has the property that $A_1(\mathbf{x})$ has exclusively negative entries. Find the resulting suggested change (that is, -1 times the learning rate times the derivative of cost with respect to the given matrix or vector) to the weights and biases for that sample.

(b) You and a friend each decide to use a neural net for a classification problem. They train their model and get poor results (that is, high training error), while you got much better results. You compare your code and find that you used the same architecture. Is it necessarily true that your friend has a bug in their code?

Solution

(a) Forward propagating this vector will result in a sequence of zero vectors. Therefore, each derivative of the form $\dot{K}.$ will be zero, and this means that the derivative of cost with respect to each weight or bias preceding the last application of $K.$ will be zero. Since the derivative of the last weight matrix is an outer product of the forward propagation vector in the preceding node and the gradient vector in the next node, it is also zero. Finally, the derivative with respect to the final bias vector is the derivative of the cost function $|\mathbf{y} - \mathbf{y}_i|^2$, which is $-2\mathbf{y}_i$ since $\mathbf{y} = 0$. Therefore, the suggested change is $2\epsilon\mathbf{y}_i$.

(b) No, because the neural net optimization problem is random and nonconvex, you typically won't get the same results on multiple runs of your own code.

Problem 33 [DR]

(a) Suppose that you have 1000 data points in $\mathbb{R}^{32}$, and you suspect that they might all lie on the same 3-dimensional affine space in that 32-dimensional space. Describe a procedure for checking whether that is the case.

(b) Describe how one might use t-SNE on a set of labeled data to build a classifier.

Solution

(a) We perform PCA (de-mean the columns and calculate the SVD), and we check whether the number of nonzero singular values is 3. If so (and only if so), then the points lie on the same 3D affine space.

(b) We apply t-SNE to reduce the dimension of the feature space, and then we use any classifier we like (k nearest neighbors, a KDE-based estimator, SVM, etc.).

Problem 34 [R]

(a) Write an R function that takes a positive integer n and returns `TRUE` if n is a multiple of 2 or 3 and `FALSE` otherwise.

(b) Write an R function `big.entries` that takes a vector as input and returns a vector which contains only the entries which exceed 100. Your function should pass the test `all(big.entries(c(0,-106,111,324,6)) == c(111,324))`

Solution

For (a), we just check the two given conditions:

R

```
is.multiple <- function(n) {n %% 2 == 0 || n %% 3 == 0}
```

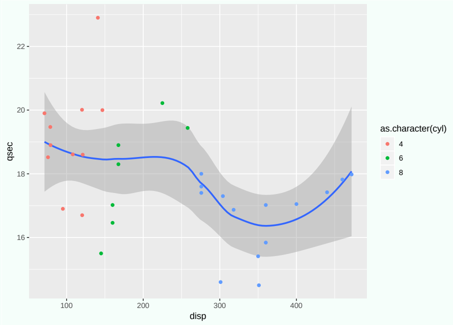
For (b), we use boolean indexing:

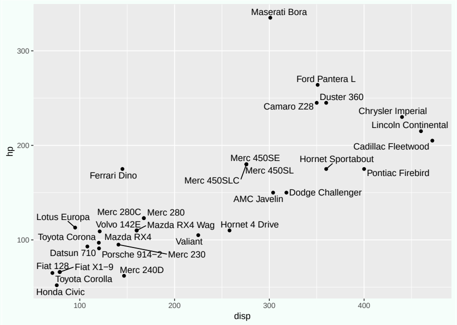
R

```
big.entries <- function(v) {v[v>100]}
```

Problem 35 [GGPLOT]

Write code to generate the figures shown. The name of the data frame is `mtcars`, and the columns include `name` (which indicates the name of the car), `disp` for engine displacement, `hp` for horsepower, `qsec` for quarter-mile time, and `cyl` for number of cylinders.





Solution

The first figure combines a point and a smooth geom:

R

```
ggplot(mtcars) +
geom_smooth(aes(x = disp, y = qsec)) +
geom_point(aes(x = disp, y = qsec, color = as.character(cyl)))
```

The second figure uses a point and a text geom:

R

```
mtcars %>%
  ggplot(aes(x = disp, y = hp, label = name)) +
  geom_text() + geom_point()
```

(Actually, the figure shown uses a variant of `geom_text` called `geom_text_repel` available in the `ggrepel` package. This package makes the labels easier to read by trying to keep them from overlapping.)

Problem 36 [DPLYR]

See the GGPLOT question for a description of the dataset `mtcars`.

(a) Write code which returns a data frame containing only the cars which have an engine displacement of at least 200 and which also has an additional column for engine displacement per cylinder.

(b) Write code which returns a data frame whose rows correspond to number of cylinders and whose columns show the minimum and maximum quarter-mile time for any car with that number of cylinders.

Solution

(a) We apply `filter` and `mutate`:

R

```
mtcars %>%
  filter(disp >= 200) %>%
  mutate(disp_per_cyl = disp/cyl)
```

(b) We group by cylinder count and summarize with min and max:

R

```
mtcars %>%
  group_by(cyl) %>%
  summarize(min.qsec = min(qsec,na.rm=true),
            max.qsec = max(qsec,na.rm=true))
```

Additional space

Additional space

Additional space

Additional space