Philosophy of Statistics: Homework 2

due on Gradescope by 11am on Tuesday April 27

Guidelines. Some questions ask you to justify your answers. For these questions, credit will be based on how well you justify your answers, not on whether your answers are correct. (There's often no consensus on the correct answers, even among statisticians.) However, that doesn't mean that anything goes: some answers will be hard to justify well. I give suggested word counts but these are just ballpark numbers. Don't sweat them too much. Collaboration is encouraged, but make sure to write up your answers by yourself and list your collaborators.

Problem 1 (20 points). We've been thinking about how to justify a preference for simplicity. But what *is* simplicity, anyway? Let's focus on functions, e.g. $f_1(x) = 2x$, $f_2(x) = 5x^2 - x + 176$, $f_3(x) = \sin(x)$, $f_4(x) = \exp\left(\sin\left(2\log_3(x^2)\right)\right)$. So our question becomes: when is f simpler than g?

A few proposals: f is simpler than g just if...

1. the simplest mechanism which computes f is simpler than the simplest mechanism which computes g
2. the representation of f is simpler than the representation of g
3. it's easier to compute an output given an input for f than for g
4. the Yu-complexity of f is lower than the Yu-complexity of g

(Yu-complexity is due to Tao Yu: take a look at their excellent forum post to see the definition.)

Discuss these proposals (300 words). Things you might like to think about:

- What if a function can be represented in multiple ways?
- What if different people find different functions easy to compute?
- How do we measure how simple a mechanism is or how simple a representation is?
- Do the proposals agree?
- Might some proposals be ill-defined in some cases?
- Might the simplicity of a function depend not just on its graph but on what the inputs and outputs represent?

This is a broad question! I'm not looking for definitive answers. Don't aim to say something about every suggestion above. Just focus, as usual, on making some clear, careful points, illustrated wherever possible with concrete examples. Your answer will be graded on that basis.

Problem 2 (15 points). The following questions are about two of the readings: Forster and Sober's *How to tell...* and ProPublica's article about COMPAS.

1. Forster and Sober say that the two steps in the standard approach to curve-fitting (choosing a family and then choosing a curve from that family) "answer to different standards". What are the standards?
2. What does 'SOS' stand for?
3. Why do Forster and Sober say that "it is overwhelmingly probable that any curve that fits the data perfectly is false"?
4. Why, according to Forster and Sober, does the fit to the data of the best-fitting curve from a family tend to improve as we move from lower to higher dimensional families?
5. ProPublica gathered data about the risk scores assigned to how many people in which county?
6. Give an example of a question from Northpointe's questionnaire.
7. What does the acronym COMPAS stand for?
8. When was ProPublica's article published?
9. In the table about halfway through ProPublica's article, titled "Prediction Fails Differently for Black Defendants", explain precisely what the number 23.5% represents.

Problem 3 (20 points). This question is designed to help you improve your grip on the Akaike Information Criterion, by applying it in a simple curve-fitting problem, similar to the example from class.

You're interested in the relationship between two properties, X and Y. So you plan an experiment: you will set $X = 1, 2, 3, 4$ and measure Y, then fit a curve to your data, using AIC.

The true curve is the curve which describes the true relationship between X and Y. Maybe Y is constant, independent of X, i.e. the true curve is some member of $F_1 := \{Y = a : a \in \mathbb{R}\}$. Or maybe Y depends linearly on X, i.e. the true curve is some member of $F_2 := \{Y = aX + b : a, b \in \mathbb{R}\}$. Or maybe the true curve is more complicated.

Your measurements, like pretty much any measurements, are subject to random measurement errors. So your measured values of Y may differ from the true values of Y. We'll suppose that the measurement errors are distributed like this:

value	-.5	-.25	0	+.25	+.5
probability	.1	.2	.4	.2	.1

This is not a realistic model of measurement error, of course, but it helps bring out the concepts more clearly.

A reminder of the Akaike Information Criterion:

> Do the experiment. Estimate the predictive accuracy of each family of curves under consideration. Choose the family which maximizes that estimate. From that family, choose the curve which best fits the data.

Some other reminders:

1. We measure the fit of a curve to data by the log of the probability of getting that data, assuming the curve is the true curve.
2. The probability of getting some data, assuming the true curve is c, is the product of the probabilities of the measurement errors.
3. We estimate the predictive accuracy of F by doing the experiment, measuring the fit of the best-fitting curve from F to the data, and subtracting k, the number of adjustable parameters in F.

Here is the data you get when you do the experiment: $(1, 1.75)$, $(2, 2.25)$, $(3, 2.25)$, $(4, 2.5)$. For example, when you set $X = 1$ you measured $Y = 1.75$.

OK, enough set-up. Let's go!

(A) Which curves from F_1 (horizontal lines) are consistent with the data?

(B) Which curves from F_2 (all lines) are consistent with the data?

(C) Which curve from F_1 best fits the data, and how well does it fit it?

(D) Which curve from F_2 best fits the data, and how well does it fit it?

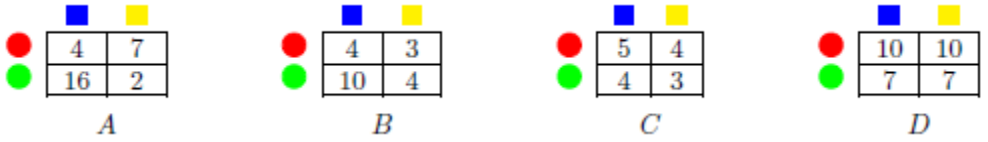
(E) What is Akaike's estimate of F_1's predictive accuracy?

(F) What is Akaike's estimate of F_2's predictive accuracy?

(G) According to AIC, which curve should we settle on?

(H) If we'd got different data, might AIC have led us to settle on a curve from the *other* family? Give an example of such data or sketch why none exists.

Problem 4 (15 points). Here is a particular crate-and-boxes problem:



Consider the following decision rule: if 'A' predict green; if 'B' predict green; if 'C' predict red; if 'D' predict green.

Work out the overall, blue and yellow confusion tables. Write the entries as whole number fractions, not as decimals.

Problem 5 (15 points). In some crate-and-boxes problems, there do exist decision rules with all four properties we talked about in class: i.e. the red prediction values, green prediction values, true red rates and true green rates are the same among blue cube boxes as among yellow cube boxes.

A crate-and-boxes problem *admits perfect prediction* if, for each label, all boxes of that label contain the same color ball. First, show that in any crate-and-boxes problem which admits perfect prediction, there exists a decision rule with all four properties.

In a crate-and-boxes problem, the *base rate* is the probability of drawing a box containing a red ball, either overall or among blue cube boxes or among yellow cube boxes. A crate has *equal base rates* if the base rate among blue cube boxes equals the base rate among yellow cube boxes. Second, show that in any crate-and-boxes problem with equal base rates, there exists a decision rule with all four properties. (To avoid some tedious edge cases, you may restrict attention to crate-and-boxes problems in which all entries in the frequency tables are non-zero.)

Third, how common do you think it is that real-life prediction problems admit perfect prediction or have equal base rates? Justify your answer. (200 words)

Problem 6 (15 points). What should we do when the consequences are uncertain? The *consequentialist* says that we should do whatever would maximize expected value. Let's try to apply that general idea in the context of choosing a decision rule in a crate-and-boxes problem.

Mac is going to draw a box at random from the crate and apply her decision rule. There are four kinds of consequences: a true positive (where we correctly predict that the ball in the box is red), a true negative (where we correctly predict that the ball in the box is green), a false positive (where we incorrectly predict that the ball in the box is red), and a false negative (where we incorrectly predict that the ball in the box is green). Let's assume we can assign numbers to these four consequences to measure how good or bad they are: TP, TN, FP, FN, respectively, where TP = TN > FP = FN. (So we're assuming that these numbers don't vary from box to box. In actual scenarios, that assumption may be wrong but let's stick with it here to make things simpler.)

The *expected value* of a decision rule is (TP multiplied by the probability of getting a true positive) plus (TN multiplied by the probability of getting a true negative) plus (FP multiplied by the probability of getting a false positive) plus (FN multiplied by the probability of getting a false negative).

Given a particular crate-and-boxes problem, the consequentialist says that Mac should use whichever decision rule maximizes expected value. Which is that?