WISDOM OF THE CROWD

FRANK SANACORY

Abstract. Large groups of diverse non-experts can match the accuracy of smaller groups of experts. In this short presentation we will go from zero to statistics and measures of variation. We will see how greater variation yields greater accuracy.

Crowds are often associated with mobs and irrationality. However, in crowds there is wisdom to be mined. This was a discovery of Francis Galton in 1906. At this fair in Plymouth England about 800 people entered a contest to estimate the weight of a ox after it had "been slaughtered and dressed." The closest estimate was the winner of the prize.

There were experts such as butchers and farmers and there were non-experts guessing the weight of the ox. The median guess was only 1 percent off of the true weight. The entire crowd guessed better than an individual expert! So people like math teachers with no experience in the weight of cattle and true experts like butchers all combined had a more accurate estimate, hence the title the wisdom of the crowd!

Where can the wisdom of the crowd be found and when do we have the madness of the mob? We will need intelligence of each individual, independence of each individual and diversity of the crowd. We will see how these interrelate.

But first, we look at some basic statistics from scratch and view two important models, categorical and linear. We will learn how to rate the effectiveness of these explanatory or predictive powers. We will learn variation, and measures of diversity and see how prediction accuracy is mathematically connected to independence and diversity.

Look at the following table.

Food	Calories
Pear	100
Cake	250
Apple	90
Banana	110
Pie	350

Table 1. Can you categorize the following to explain some of the difference in calories? What categories would you choose?

Hours	Grade (letter)	Grade (GPA)
2	A	4
3	A	4
1	C	2
2	B	3

TABLE 2. If a student studies 2.5 hours what would you predict the grade to be? How about a student who studies for 1.5 hours?

Food	Calories	Diff ($s_i - c$)	Diff squared ($(s_i - c)^2$)
Pear	100	100-180 = -80	$(-80)^2 = 6400$
Cake	250	70	4900
Apple	90	-90	8100
Banana	110	-70	4900
Pie	350	170	28900

TABLE 3. New caption

1. Categorical Models

Let's look at our foods table. What did you see as a an explanatory difference? Possibly you said some of the foods are desserts and some are fruits. We can divide the foods into two categories and then measure how much variation this categorization explains. So we will measure

To compute variation we will

(1) Compute the average of the statistic (c).
(2) Compute how much the statistic and the average differ? To compute this difference we subtract.
(3) Square that difference. Why do you think we square it?
(4) the variation is the sum of those squared differences.

For our average I get

$$c = \frac{100 + 250 + 90 + 110 + 350}{5} = \frac{900}{5} = 180.$$

Summing the last column get $6400 + 4900 + 8100 + 4900 + 28900 = 5320$.

Now we break the data into the two categories and repeat. FOr the dessert category mean I get

So I get the variation in the fruit category is 200 and the variation in the dessert category is 5000. Thus we have a variation of 5200. [1]

So we explained 53200-5200 = 48000 of the variation. Or we explained $48000/53200 \approx 90.2\%$ of the variation. We call this R^2.

$R^2 = 1 - \frac{\text{model variation}}{\text{total variation}}$

[1] What are the units of variation?

Food	Calories	Diff $(s_i - c)$	Diff squared $(s_i - c)^2$
Pear	100	100-100 $= 0$	$0^2 = 0$
Apple	90	-10	100
Banana	110	10	100
			200

Table 4. New caption

Food	Calories	Diff $(s_i - c)$	Diff squared $(s_i - c)^2$
Cake	250	250-300=+50	2500
Pie	350	50	2500
			5000

Table 5. New caption

When R^2 is near 1 (or near 100%), we have explained nearly all of the variation and we have a good model. When R^2 is near zero (0%) we have explained almost no variation and our model is not good.

very little variation explained
much variation variation explained
0
1

Figure 1. When R^2 is near zero the model explains very little variation, and when the R^2 is near one the model explains much of the variation.

You do. I really want to watch all of my dvds for the TV show Friends on new TV. So I went shopping and listed all of the prices I found. I found the following prices (see Table 6). Using two categories, Sony and Hisense, how much variation can you explain?

2. Linear Models

We will use the equation of the line we are all familiar with $Y = mX + b$. Recall the data for the grades.

So we need to do the same as before, compute total variation and compute the model's variation. To compute the total varation the computation is identical. We will use grades in place of calories and compute, the mean, the differences and the square differences. Then

TV	Price
Sony KD50X80J 50”	\$1,200
SONY XR65X95J Bravia XR X95J 65”	\$2,000
Sony X95J 75 Inch TV	\$3,000
Hisense 55-Inch Class H9 Quantum Series	\$700
Hisense 65-Inch Class H8 Quantum Series	\$1,000
Hisense 75U9DG ULED 75 Inch	\$2,500

Table 6. Prices are from Amazon.com with some rounding.

Hours	Grade (letter)	Grade (GPA)
2	A	4
3	A	4
1	C	2
2	B	3

Table 7. New Caption

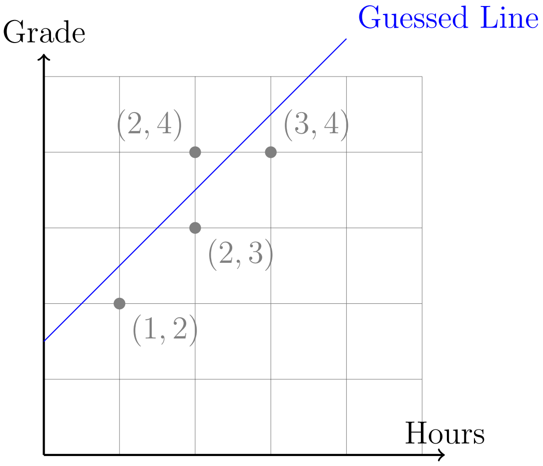

Figure 2. The guessed line is $y = x + 1.5$. Is it ideal?

Hours	Grade (letter)	Grade (GPA)	Diff	Diff Squared
2	A	4	4-3.25 = 0.75	$(0.75)^2 = 0.5625$
3	A	4	0.75	0.5625
1	C	2	-1.25	1.5625
2	B	3	-0.25	0.0625

Table 8. Summing the Diff Squared we get 2.75.

we sum the square differences (as in Table 8). I get the mean of the GPA grade as 3.25.

To compute the ,model error we need to first compute the prediction for each student. The first student has studied for 2 hours and received

Hours	Grade (letter)	Grade (GPA)	model ($y = x + 1.5$)	Diff (model-grade)	Diff Squared
2	A	4	3.5	4-3.5 = 0.50	$(0.50)^2 = 0.25$
3	A	4	4.5	-0.50	0.25
1	C	2	2.5	-0.50	0.25
2	B	3	3.5	-0.50	0.25

Table 9. Summing the Diff Squared we get 1.

a grade of 4. The model predicts a grade of

$$\begin{aligned} y &= x + 1.5 \\ \text{model predicted grade} &= hours + 1.5 \\ \text{model predicted grade} &= 2 + 1.5 = 3.5. \end{aligned}$$

So the first student received a grade of 4, and our model predicted a grade of 3.5. To compute the difference for the model we subtract actual grade minus model prediction for the grade. That is,

$$\text{Diff} = \text{Actual Grade} - \text{Model Prediction} = 4 - 3.5 = 0.5.$$

Then we square the difference as before. See Table 9.

Our Total Variation is 2.75 and our model variation is 1. So we have an R^2 of

$$R^2 = 1 - \frac{1}{2.75} = 63.6\%.$$

The model explains some of the variation.

You do. I am still searching for a new TV (now I want to watch the Umbrella Academy). Using the prices from Table 6, guess a linear model with the x variable as the inches and the y variable as the price. Calculate the variation explained by your model?

You do. How can we improve our TV explanation model? What do you think.

3. Diversity Prediction Theorem

So far we have been looking at basic statistics and ignoring the orginal problem of a crowd with guesses. We will look at the crowd and those guesses and use the computation of variance to help us understand. So we have a crowd with some guesses. We compute the mean (or Galton used the median) to get a better guess ad to the true number.

The Diversity Prediction Theorem roughly says

- More accurate independent guesses $\implies$ more accurate crowd average
- More diversity in the crowd $\implies$ more accurate crowd average

Let's look at an example. Let's say we have three guesses, 10, 16 and 25, for an item which has true value of 18. We will use the notation of

- c for crowd guess, so c is the mean.
- s_i for the individual guesses.
- θ is the true value.

Let's compute three things,

(1) average individual error,

$$\frac{1}{n}\sum_{k=1}^{n}(s_i-\theta)^2$$

(2) crowd error, and

$$(c-\theta)^2$$

(3) diversity (the average variation of the crowd).

$$\frac{1}{n}\sum_{k=1}^{n}(s_i-c)^2$$

average individual error

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}(s_i-\theta)^2 &= \frac{1}{3}\sum_{k=1}^{n}(s_i-18)^2\\ &= \frac{1}{3}[(10-18)^2+(16-18)^2+(25-18)^2]\\ &= \frac{1}{3}[64+4+49]=39\end{aligned}$$

crowd error

$$(c-\theta)^2=(17-18)^2=1$$

diversity (the average variation of the crowd).

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}(s_i-c)^2 &= \frac{1}{3}\sum_{k=1}^{n}(s_i-17)^2\\ &= \frac{1}{3}[(10-17)^2+(16-17)^2+(25-17)^2]\\ &= \frac{1}{3}[49+1+64]=38\end{aligned}$$

Notice we get

(1) average individual error = 39,
(2) crowd error = 1, and
(3) diversity (the average variation of the crowd) = 38.

Does this suggest a formula?
The Diversity Prediction Theorem states

$$\text{Crowd Error} + \text{Diversity} = \text{Average Indivudual Error}$$

$$(c-\theta)^2 + \frac{1}{n}\sum_{k=1}^{n}(s_i - c)^2 = \frac{1}{n}\sum_{k=1}^{n}(s_i-\theta)^2$$

4. Extras

The best linear model is about $price = -3030 + 74 * (inches)$.

We can put the two models together by coding Sony as a 1 and Hisense as a zero. We get the following "best" model.

$$price = -3674 + 797 * (brand) + 78 * (inches).$$

So for aexample a 70 inch sony would cost

$$price = -3674 + 797 * (1) + 78 * (70) = 2583.$$

And a 68 inch Hisense would cost

$$price = -3674 + 797 * (0) + 78 * (68) = 1630.$$

What is the R^2 for this multiple Model?

References

[1] Galton, Francis. "Vox populi (the wisdom of crowds)." *Nature* 75.7 (1907): 450-451.

[2] Page, Scott E. *The model thinker: What you need to know to make data work for you.* Basic Books, 2018.

[3] Page, Scott. *The diversity bonus.* Princeton University Press, 2019.

Department of Mathematics & CIS, Associate Professror of Mathematics, Chair

Email address: sanacoryf@oldwestbury.edu

URL: www.sanacory.net