CMPLXSYS/POLSCI 391 — Winter 2022 — Discussion
Diversity Prediction Theorem & Categorization Error
Bhaskar Kumawat

Legend

Text in red are questions that need to be submitted as a part of the in-class discussion before midnight today.

Text in blue are questions that need to be submitted as a part of the home due before next lab.

Diversity Prediction Theorem (DPT)

Statement

The error from using multiple models together is less than the average error of these models, by an amount exactly equal to the diversity of the model predictions.

Formal Equation

- M_i = Prediction of model i
- $\bar{M}$ = Average of prediction of all models = $\frac{\sum_{i=1}^{N} M_i}{N}$
- V = True value

$$\underbrace{\left(\bar{M} - V\right)^2}_{\text{Many-model error}} = \underbrace{\sum_{i=1}^{N} \frac{\left(M_i - V\right)^2}{N}}_{\text{Average-model error}} - \underbrace{\sum_{i=1}^{N} \frac{\left(M_i - \bar{M}\right)^2}{N}}_{\text{Prediction Diversity}} \tag{1}$$

Let's break it down

Let us say you estimate the outcome of a process in the world to be a value X (For example, X can be the estimated value of the number of people that vote Democrat before the elections). After the process, you find that the actual outcome is V (true value). The error in your measurement is then given by,

$$\text{Error} = (X - V)^2 \tag{2}$$

The difference is squared because whether you overshoot or undershoot the real value doesn't really matter when you are calculating the magnitude of the error.

Many-model error: The Many-model error is the error assuming your estimate is the average of what you get from your different models (i.e., $X = \bar{M}$)

Average-model error: The average-model error is the average of all the errors when you use individual model outcomes as your estimators.

Prediction diversity: The prediction diversity calculates how different the individual model outcomes are from the average.

An Example

You are a climate scientist working on estimating the mean temperature of earth's surface in 2050. You come up with multiple estimates of the temperature using different climate models as follow.

$M_1 = 4\,°C$
$M_2 = 6\,°C$
$M_3 = 10\,°C$
$M_4 = 8\,°C$
$M_5 = 7.5\,°C$
$M_6 = 3\,°C$
$M_7 = 10\,°C$
$M_8 = 11\,°C$

1. What is the average of all of these model estimates? ($\bar{M}$)

2. What is the prediction diversity?

3. Let us assume the scientist lives till 2050 and finds out that the actual temperature recorded was 8 $°C$. What is the many-model error if the scientist had used $\bar{M}$ as the estimator?

4. Another group of 8 scientists decided to not pool there results and instead went each with their own model. What is the average model error of this group?

5. By what value did the single scientist improve the estimate (over the group of 8) by taking an average of all the models?

Universal biases propagate (Homework)

6. Let us say that all the models that scientists used (M_1 to M_8) had a positive bias and actually estimated the temperature to be 1 $°C$ higher. Recalculate the three terms of the DPT equation.

7. Which of the terms remained the same? Which of the terms changed? How does it limit our ability to predict things with zero error?

Model Error Decomposition Theorem

Statement

The total error of a model is the sum of errors from categorizing a data-point and the error from using a particular model for that category.

Formal Equation

- n = Number of categories
- x = A data-point you want to estimate
- V_i = True value of mean for category i
- $V(x)$ = True value of the data point x
- M_i = Model value for category i
- $M(x)$ = Model value for data-point
- $|C_i|$ = Number of data-points in the category

$$\underbrace{\sum_{x\in X}(M(x)-V(x))^2}_{\text{Model Error}} = \underbrace{\sum_{i=1}^{n}\sum_{x\in S_i}(V(x)-V_i)^2}_{\text{Categorization error/loss}} + \underbrace{\sum_{i=1}^{n}|C_i|\,(M_i-V_i)^2}_{\text{Valuation Error}} \tag{3}$$

An Example

Consider the following data about the yearly income of certain employees in a company and the respective hours they work.

Employee#	Yearly income (1000$)	Hours
1	60	40
2	80	50
3	110	50
4	150	30
5	200	20
6	300	10
7	500	20

1. Divide the employees into two categories, those who earn less than 100,000$ a year (A) and those who earn more than 100,000$ a year (B). What are the average value of **hours** for these two categories? (These are your V_i's)

2. When we categorize, we are basically relabelling the hour values for each individual to be equal to the average of its category. What is the total error when we do this relabelling for each category? (This is your $\sum_{x \in S_i}(V(x) - V_i)^2$)

3. What is the sum of categorizing errors for the two categories? (This is your total categorization error/loss!)

4. Using a particular model (M) of how income affects productivity, you find out the following estimates of the productivity of each category (M_is).

- Category A (low income): Hours worked = 50 hours
- Category B (high income): Hours worked = 20 hours

What is the error for each category if your estimate is the value from the model (M_i) but the true value is the mean for each category (V_i)? (i.e., $(M_i - V_i)^2$ for both categories)

5. This error repeats for each employee in these categories. So the total valuation error for each category is (No. of employees in category) $\times (M_i - V_i)^2$. What is the sum of total valuation error for the two categories? (This is your total valuation error from using the model!)

6. The model assigns its predicted productivity to each employee x depending on what it predicted for the categories. That is, $M(x) = 50$ if x is in category A and $M(x) = 20$ if x is in category B. We already know the true values of productivity ($V(x)$) for each employee (first table). Calculate the total model error. ($\sum_x (M(x) - V(x))^2$)

7. Verify that the total model error is the sum of the categorization error and the valuation error. This means that how well a model predicts something (accuracy) depends on how well it categorizes things and reduces categorization errors. At the same time, categorizing everything into separate categories (maximizing accuracy) decreases predictive power because a new data-point might not have the same value as any of the other points in our dataset (and thus might not belong to any of the categories)!

Discussion Feedback (anonymous; though you'll need to login to your UM google account to prevent unauthorized access): `https://forms.gle/WyP7o66k6hnozcHA8`