Thursday Thunder: Week 7,Learning 7: Simpson's Paradox- Averages Gone Wild

Manish Tripathi Mon, Feb 29, 2016 at 2:55 PM
To: PDS-ALL

Hi

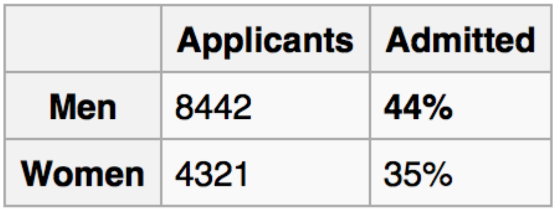
Last week Esther told me in an informal conversation that how the direction of one of the feature weights in the model is changing sign when certain features are removed or added. So that motivated me to write a small post on the same.

What is Simpson's Paradox

Well, it is an effect which one sees in discrete data in which the association between two variables might get completely reversed if you look at the data in aggregation compared to looking at data in dis-aggregation.

What does that supposed to mean

Let me explain by an example. Here is the famous UC Berkeley example in which the University was sued for gender discrimination in their admissions acceptance rate. Here is the aggregated data from the University.

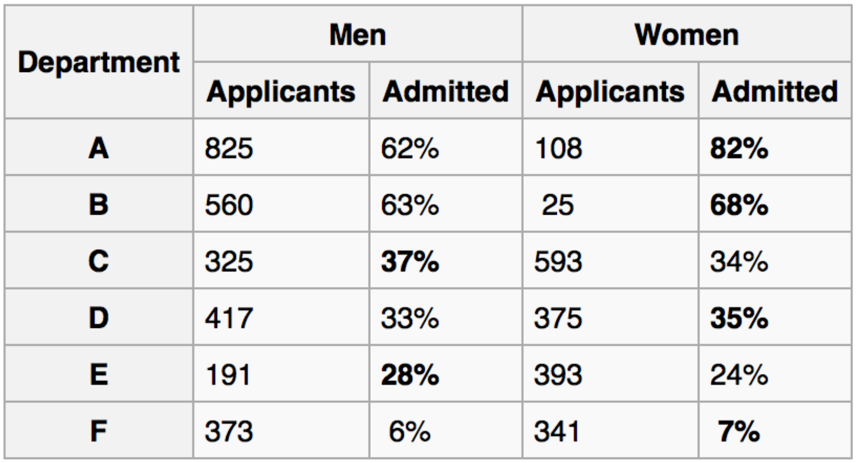


	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

Based on this data, people sued the University saying they are biased against Women.

It does look like they admitted more Men.

Yes, it does. But take a look at this now. UC Berkeley went back to do further investigation and found the following data by department -wise acceptance rate.



Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

Well now the picture is totally reversed. It looks like Women were actually getting more admission by department than men.

Am Confused. Which one is correct. What is going on?

Ha ha. This is what is known as Simpson's Paradox. If you look at the data and try taking a decision based on aggregate data, you might find a completely different association between variables and hence a opposite decision than when you look at the data at dis-aggregation level.

The reason that happens is because when you aggregate the data, there is a confounder or hidden lurking variable called Department which changes the behavior of admission amongst gender. Apparently, it was found that lot of Women were applying for departments where admission rate is low, while majority of men were applying for departments were acceptance rate is usually high. Look at the absolute numbers in the above table.

Mathematically , Simpson's Paradox is a property of unreduced fractions.

This is where Averages go wild and tend to mislead in decision process. ***Averages lie a lot and hide lot of information than what you see***

Ok. How does it all relate to Machine Learning.

Ok. So now you know that partitioning data can lead to different conclusion than looking at aggregate data, due to Simpson's Paradox. If you now can understand that adding features into your model especially categorical features is akin to partitioning your data by those categorical variables ,then it would tend to give an association between a feature and dependent variable completely opposite than what you might observe without that categorical feature added. Because now you are dis-aggregating the data by that categorical feature compared to a scenario where you were aggregating it across the categories of that categorical feature.

Hmm. So how do you take care of that.

One of the suggested approach is to use a graphical model , especially a DAG (Directed Acyclic Graph also known as Bayesian Network). That would give you the direction of the relationship between the variables and would also give a sense if there is a possibility of reversal of association (change in the direction of arrow) between two variables.