CS 6604: Data Mining Fall 2007

Lecture 13 — October 1, 2007

Lecture: Naren Ramakrishnan *Scribe: Naveed Massjouni*

1 Overview

In the last lecture we discussed inverse resolution and elephants. In this lecture we will discuss θ-subsumption. We begin by reviewing inverse resolution. Then we will learn about θ-subsumption and its relationship to inverse resolution. Finally, we will discuss the delicate dance between bias and variance.

2 Inverse Resolution

Recall that resolution is an inference rule. Consider the following:

$$\frac{\neg p \vee q,\ \neg q \vee r}{\neg p \vee r}$$

Resolution is taking the information in the numerator and concluding the denominator. It goes from more general to more specific. Inverse resolution goes in the opposite direction, from specific to general. Given $\neg p \vee r$ and $\neg p \vee q$, it would find $\neg q \vee r$.

3 θ-subsumption

Clause C_1 θ-subsumes clause C_2 if there exists a substitution θ s.t. $C_1\theta \subseteq C_2$.

Example 1:
$C_1 = p(x, y)$
$C_2 = p(John, Mary)$
$\theta = \{x/John, y/Mary\}$

Example 2:
$C_1 = a(X_0, X_1), a(X_0, X_2), a(X_0, X_3), b(X_1, X_2), b(X_1, X_3)$
$C_2 = a(Y, Y_1), a(Y, Y_2), b(Y_1, Y_2)$
$\theta = \{X_0/Y, X_1/Y_1, X_2/Y_2, X_3/Y_2\}$

The following holds true if if C_1 θ-subsumes C_2:

- $C_1 \models C_2$.
- C_1 is at least as general as C_2.

4 Inverse resolution vs θ-subsumption

Inverse Resolution	**θ-subsumption**
bottom-up	top-down
semantic	syntactic
implies θ-subsumption	does NOT imply inverse resolution

5 Entailment and Generality

Like θ-subsumption, the entailment relationship is one way to relate general to specific formulas. Lets consider the entailment relationship: $p \models p \vee q$.

There are two models of the left side of the above equation:

p	**q**
1	0
1	1

There are three models of the right side:

p	**q**
0	1
1	0
1	1

Observe that the first set of models is a subset of the second set of models. That is why we say p entails $p \vee q$. If models of m_1 are a subset of the models of m_2, then m_1 entails m_2.

As to which one is more general, we say that p is more general than $p \vee q$ because we logically concluded the latter from the former. Deduction proceeds from the general to the specific. Therefore, this is one definition of "general": X is more general than Y if X entails Y.

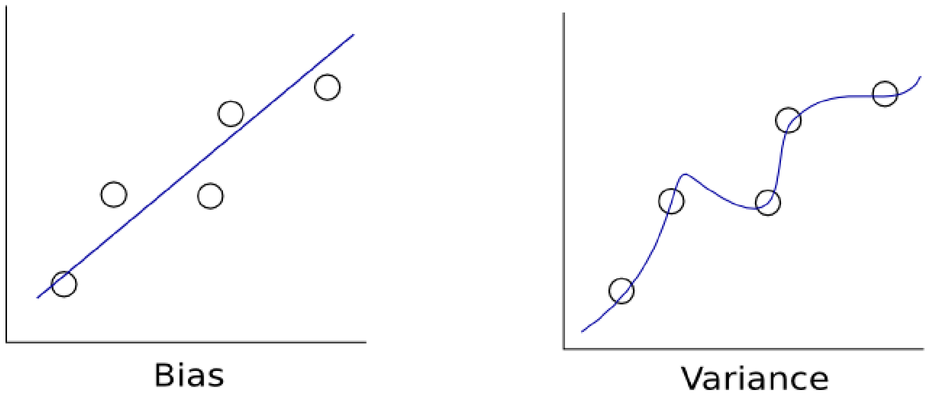
Another way to think about this is to observe that $p \vee q$ is shorthand for $\neg q \rightarrow p$. Now compare:

$\neg q \rightarrow p$ (i.e., $p \vee q$)
with
p

The former says "you can conclude p if (not q) is true". The latter says "you can conclude p", period. No conditions. Hence it is more general.

6 Bias vs Variance

There is a delicate balance that data miners need to consider with regards to bias and variance.



In the above figure there are two models of the same data. The model on the left has high bias and low variance. The model on the right has high variance and low bias. Choosing a model with high bias will help to filter out noise, but it may not be sufficient if the correct model is complex. Choosing a model with high variance would allow you to model complex data, but your model might be overfitting the data and may not generalize.