Probabilitistic Graphical Models for Detecting Underwriting Fraud

Mick Cooney
michael.cooney@applied.ai

11 July 2016

How to Build a Model with No Data and No Domain Knowledge...

Structure of Talk

Medical Non-disclosure

Bayesian Networks

Building the Model

Conclusions

Medical Non-disclosure

SSQ Financial Group

REQUEST FOR OPTIONAL LIFE INSURANCE

PLEASE COMPLETE THIS FORM IN BLOCK LETTERS USING INK

A. EMPLOYER INFORMATION

Policy Holder Name:	SSQ Group #:
Division Name:	Certificate #:

B. PARTICIPANT INFORMATION

Last Name:	First Name:	S.I.N.:
Mailing Address: (including postal code)		
Telephone: Home	Work	Language Preference: ☐ English ☐ French
Gender: ☐ M ☐ F	Date of Birth: D M Y	Salary: $

C. REQUEST FOR OPTIONAL LIFE INSURANCE COVERAGE

IMPORTANT: Optional Life Insurance units of $10,000 are only available to plans that currently offer this benefit.

Participant: ***(Please check N/A if request is only for spouse)***

Spouse:

Participant: Current amount of coverage (in force)	Participant: Additional amount of coverage (requested)	Spouse: Current amount of coverage (in force)	Spouse: Additional amount of coverage (requested)
☐ None ☐ 1x salary ☐ 2x salary ☐ 3x salary ______ units of $10,000	☐ N/A ☐ 1x salary ☐ 2x salary ☐ 3x salary ______ units of $10,000	☐ None ☐ 25% ______ units of $10,000	☐ 25% ☐ 50% ______ units of $10,000

Spouse:	Last Name:	First Name:
	Gender: ☐ M ☐ F	Date of Birth: D M Y

D. SMOKING HABITS

Participant: Non-Smoker ☐ Smoker ☐ **Spouse:** Non-Smoker ☐ Smoker ☐

"I declare that I do not smoke and have not smoked any tobacco products such as cigarettes, cigars, cigarillos or pipes, or any drugs during the past 12 months. This statement is an affirmative guarantee on my part." It is understood that the insurer may periodically require confirmation of non-smoker status. The participant must be in a position to meet the requirements then in force and return the confirmation within 30 days of the request, failing which the participant shall lose non-smoker status and the associated premium reduction shall cease to apply as of the date of the insurer's request. "I also acknowledge that a false or incomplete statement may cause the coverage to be null and void."

Participant: ______________________ Spouse: ______________________

Problems

Data sparse / missing

Partially missing output variable

Low base-rate problem

Semi-supervised learning

Fraud Detection

Full automation difficult!

Create filter instead — triage cases

Build a Model

We want a model which, given the data observed in the policy application, allows us to estimate the probability of a subsequent medical exam changing the underwriting decision on the policy.

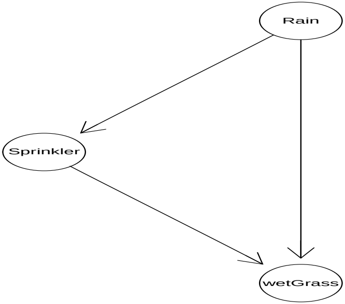
The model should incorporate our assumptions of the process and be as simple as possible.

Is the Juice Worth the Squeeze?

Probabilistic Graphical Model?

Bayesian Networks

PGM with *directed, acyclic graph* (DAG):



Variables: (R)aining, (S)prinkler, wet(G)rass

Conditional Probability Tables (CPTs)

Some Questions

What is the probability of the grass being wet?

```
querygrain(sprinkler_grain, nodes = 'wetGrass')$wetGrass

## wetGrass
##  yes   no
## 0.44 0.56
```

If the grass is wet, what is the probability that it is raining?

```
querygrain(sprinkler_grain
          ,evidence = list(wetGrass = 'yes')
          ,nodes = 'Rain')$Rain

## Rain
##  yes   no
## 0.41 0.59
```

Getting Started

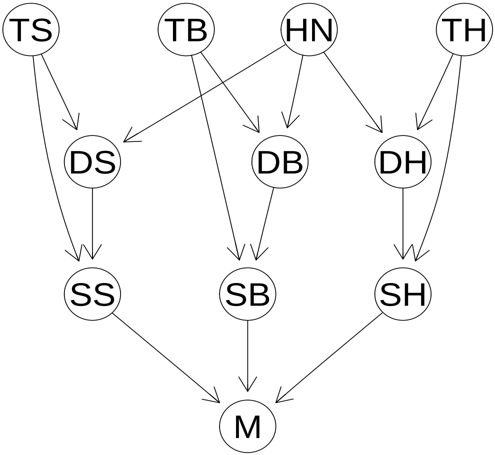
Conditions:

- **(S)moker**: Smoker, Quitter, Non-smoker
- **(B)MI**: Normal, Overweight, Obese
- **Family (H)istory**: None, HeartDisease

Aspects:

- **T**: True state
- **D**: Declared state
- **S**: Seriousness of condition's impact on decision

Medical Exam Network



- *HN*: Honesty
- *TS*: True Smoking
- *DS*: Decl Smoking
- *SS*: Serious Smoking
- *TB*: True BMI
- *DB*: Decl BMI
- *SB*: Serious BMI
- *TH*: True History
- *DH*: Decl History
- *SH*: Serious History
- *M*: Medical Chance

Assess

What is the unconditional probability of a medical exam finding something?

```
querygrain(underwriting_grain, nodes = 'M')$M

## M
##   Medical NoMedical
##      0.18      0.82
```

Too high?

Probably flawed

Assess the Model

Declares a clean bill of health (DS = Nonsmoker, DB = Normal, DH = None)?

```
querygrain(underwriting_grain, nodes = 'M'
          ,evidence = list(DS = 'Nonsmoker'
                          ,DB = 'Normal'
                          ,DH = 'None'))$M

## M
##   Medical NoMedical
##      0.15      0.85
```

Expanding the Model

Guessed CPTs — use data?

CPTs assist this - subsets of variables available

Bootstrap to validate?

Add states/levels to variables – `HeartDisease`?

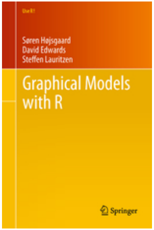
Add variables: Family History, Medical Exams, Honesty?

Conclusions

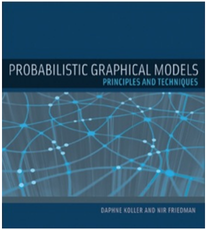
- Classification very difficult
- Highly speculative – nowhere near production-ready
- Use as filter – no automation
- Outputs often counter-intuitive
- Work unfinished - lots more avenues to explore

Other areas: Claims fraud, product recommendations, regulatory issues

Further Resources



"Graphical Models with R"
Søren Højsgaard.



"Probabalistic Graphical Models: Principles and Techniques"
Koller and Friedman

Package Vignettes: `gRain` and `gRbase`

Coursera: Probabilistic Graphical Models `https://www.coursera.org/course/pgm`

Get In Touch

Mick Cooney
michael.cooney@applied.ai

Slides and code available on GitHub:
https://www.github.com/kaybenleroll/dublin_r_workshops

Blogpost Series:
http://blog.applied.ai/probabilistic-graphical-models-for-fraud-detection-part-1
http://blog.applied.ai/probabilistic-graphical-models-for-fraud-detection-part-2
http://blog.applied.ai/probabilistic-graphical-models-for-fraud-detection-part-3