# From Linguistic Rules to Machine Learning

Cohan Sujay Carlos
Aiaioo Labs
Bangalore, India
cohan@aiaioo.com

# What is a Classifier?

A machine learning tool used to apply a label to data.

# Classification used in Text Categorization

## Politics

The UN Security Council adopts its first clear condemnation of Syria for its continuing crackdown on protests, as the army continues its advance into Hama.

## Sports

Warwickshire's Clarke equalled the first-class record of seven catches for an outfielder in an innings but Lancashire took control on day three.

# How to Build a Classifier for Text Categorization

How do you tell which label (Politics/Sports) is suitable?

The UN Security Council adopts its first clear condemnation of Syria for its continuing crackdown on protests, as the army continues its advance into Hama.

Warwickshire's Clarke equalled the first-class record of seven catches for an outfielder in an innings but Lancashire took control on day three.

In this case **it's only words**

that you will need!

# Classification used in Text Categorization

See the words?

The UN Security Council adopts its first clear condemnation of Syria for its continuing crackdown on protests, as the army continues its advance into Hama.

Warwickshire's Clarke equalled the first-class record of seven catches for an outfielder in an innings but Lancashire took control on day three.

# Rule-Based Text Categorization

Gazetteers (word lists)

UN Security Council
Adopts
Condemnation
Syria
Crackdown
Protests
Army
Hama

Warwickshire
Clarke
First-class
Record
Catches
Outfielder
Innings
Lancashire

So you can just use word lists for classification?
Yeah, but they won't work very well.
**Can you see why word lists alone won't work very well?**

# Rule-Based to Naïve Bayesian

**How can you go:**

from the starting point (word lists)

to a really cool classification algorithm

All you need is **weights**!

# Rule-Based with Weights

Let's improve the gazetteers with weights

| Politics | | Sports | |
|---|---|---|---|
| UN | 1.0 | Warwickshire | 0.3 |
| Adopts | 0.1 | Clarke | 0.1 |
| Condemnation | 0.2 | First-class | 0.6 |
| Syria | 0.3 | Record | 0.3 |
| Crackdown | 0.8 | Catches | 0.6 |
| Protests | 1.0 | Outfielder | 1.0 |
| Army | 0.8 | Innings | 0.9 |
| Hama | 1.0 | Lancashire | 0.5 |

These weights are nothing but P(Category|Word).

# Rule-Based with Weights

Let's improve the gazetteers with weights

## Politics

| | |
|---|---|
| UN | 1.0 |
| Adopts | 0.1 |
| Condemnation | 0.2 |
| Syria | 0.3 |
| Crackdown | 0.8 |
| Protests | 1.0 |
| Army | 0.8 |
| Hama | 1.0 |

**P(Politics|Word)**

## Sports

| | |
|---|---|
| Warwickshire | 0.3 |
| Clarke | 0.1 |
| First-class | 0.6 |
| Record | 0.3 |
| Catches | 0.6 |
| Outfielder | 1.0 |
| Innings | 0.9 |
| Lancashire | 0.5 |

**P(Sports|Word)**

# Rule-Based with Weights

**Politics**

| | | |
|---|---|---|
| UN | 1.0 | P(Politics\|UN) |
| Adopts | 0.1 | P(Politics\|Adopts) |
| Condemnation | 0.2 | P(Politics\|Condemnation) |
| Syria | 0.3 | P(Politics\|Syria) |
| Crackdown | 0.8 | P(Politics\|Crackdown) |
| Protests | 1.0 | P(Politics\|Protests) |
| Army | 0.8 | P(Politics\|Army) |
| Hama | 1.0 | P(Politics\|Hama) |

# Rule-Based with Weights

## Politics

| | | |
|---|---|---|
| UN | 1.0 | P(Politics \| “UN”) |
| Adopts | 0.1 | P(Politics \| “Adopts”) |

How can you learn these probabilities automatically?

# Rule-Based with Weights

## Politics

| | | |
|---|---|---|
| UN | 1.0 | P(Politics \| "UN") |
| Adopts | 0.1 | P(Politics \| "Adopts") |

How can you learn these probabilities automatically?

## Estimation

P(Politics | "UN") = 20/20

**Statistically not a very accurate estimator - denominator is small.**

# Rule-Based with Weights

## Politics

| | | |
|---|---|---|
| UN | 1.0 | P(Politics \| "UN") |
| Adopts | 0.1 | P(Politics \| "Adopts") |

How can you learn these probabilities automatically?

## Instead you Estimate

P("UN" | Politics) = 20/40000

{ C("UN" in politics) / C(all words in category politics) }

**Statistically this is a better estimator**

# Rule-Based with Weights

**Politics**

| | | |
|---|---|---|
| UN | 1.0 | P(Politics \| "UN") |
| Adopts | 0.1 | P(Politics \| "Adopts") |

How can you learn these probabilities automatically?

A Naïve Bayesian classifier uses a P(Politics | "UN") estimate calculated from P("UN"|Politics).

**That's so cool! Time to learn how to do that!**

# Use Bayesian Inversion!

In other words, we are looking to turn

$P(F|E)$ into $P(E|F)$.

There is an equation to do this :

$P(E|F) = P(F|E) * P(E) / P(F)$ [Bayesian Inversion]

# So finally … you have …

**Politics**

| | | |
|---|---|---|
| UN | 1.0 | P(“UN”|Politics)* P(Politics)/P(“UN”) |
| Adopts | 0.1 | P(“Adopts”|Politics)*P(Politics)/P(“Adopts”) |

**That was easy wasn’t it?!**

These don’t have to be only words. They can be ANY sort of feature (word pairs, syntax).

# You have just Learnt How to Build A Naïve Bayesian Classifier Starting from Linguistic Rules (Word Lists)

Cohan Sujay Carlos
Aiaioo Labs
Bangalore, India
cohan@aiaioo.com