# Naive Bayes

Samuel Clark

In collaboration with the WHO VA Reference Group
Columbus, Ohio
November, 2018

# Outline

# Background

Naive Bayes is at the heart of three available algorithms

- Naive Bayes Classifier for VA from Toronto group [3]
- InterVA from Peter Byass [1]
- InSilicoVA from our group [2]

# Background

Naive Bayes is at the heart of three available algorithms

- Naive Bayes Classifier for VA from Toronto group [3]
- InterVA from Peter Byass [1]
- InSilicoVA from our group [2]

With a simplifying assumption, naive Bayes provides an analytical relationship for the probability of something being true in a specific circumstance when something else is true

# Background

Naive Bayes is at the heart of three available algorithms

- Naive Bayes Classifier for VA from Toronto group [3]
- InterVA from Peter Byass [1]
- InSilicoVA from our group [2]

With a simplifying assumption, naive Bayes provides an analytical relationship for the probability of something being true in a specific circumstance when something else is true

Formally, this can be expressed as a *conditional probability*: the probability of the outcome given a set of conditions, or

$$\Pr(\text{outcome}|\text{conditions})$$

# Notation

- $C$ causes of death indexed by $c$

# Notation

- $C$ causes of death indexed by $c$
- $S$ binary-coded symptoms indexed by $s$

# Notation

- $C$ causes of death indexed by $c$
- $S$ binary-coded symptoms indexed by $s$
- $\mathbf{s}$: for a given death, $S$-element vector with one binary value (0=absent, 1=present) for each symptom

# Bayes' Rule

Using Bayes' Rule for conditional probabilities, for a single death the joint probability of a specific cause $c$ and a specific vector of symptoms $\mathbf{s}$ is

$$\Pr(c, \mathbf{s}) = \Pr(c|\mathbf{s}) \Pr(\mathbf{s}) = \Pr(\mathbf{s}|c) \Pr(c) \ , \tag{1}$$

# Bayes' Rule

Using Bayes' Rule for conditional probabilities, for a single death the joint probability of a specific cause $c$ and a specific vector of symptoms $\mathbf{s}$ is

$$\Pr(c, \mathbf{s}) = \Pr(c|\mathbf{s}) \Pr(\mathbf{s}) = \Pr(\mathbf{s}|c) \Pr(c) \ , \tag{1}$$

and the conditional probability of a cause $c$ given $\mathbf{s}$ is

$$\Pr(c|\mathbf{s}) = \frac{\Pr(c)\Pr(\mathbf{s}|c)}{\Pr(\mathbf{s})} \ . \tag{2}$$

# Bayes' Rule

Using Bayes' Rule for conditional probabilities, for a single death the joint probability of a specific cause $c$ and a specific vector of symptoms $\mathbf{s}$ is

$$\Pr(c, \mathbf{s}) = \Pr(c|\mathbf{s})\Pr(\mathbf{s}) = \Pr(\mathbf{s}|c)\Pr(c) \ , \tag{1}$$

and the conditional probability of a cause $c$ given $\mathbf{s}$ is

$$\Pr(c|\mathbf{s}) = \frac{\Pr(c)\Pr(\mathbf{s}|c)}{\Pr(\mathbf{s})} \ . \tag{2}$$

**Our job is to define the factors in the RHS of Eq 2.**

# Conditional Independence

**Making the assumption that observed symptoms are independent of each other for each cause of death**, the probability of a given vector **s** of symptom indicators if the cause of death is $c$ is

$$\Pr(\mathbf{s}|c) = \prod_{s} \Pr(s|c)^{s}(1 - \Pr(s|c))^{1-s} \,. \tag{3}$$

# Conditional Independence

**Making the assumption that observed symptoms are independent of each other for each cause of death**, the probability of a given vector **s** of symptom indicators if the cause of death is $c$ is

$$\Pr(\mathbf{s}|c) = \prod_{s} \Pr(s|c)^{s}(1 - \Pr(s|c))^{1-s} \ . \tag{3}$$

This is **not** realistic, but it greatly simplifies things.

# Probability of a Vector of Symptom Indicators

The probability of a given vector of symptoms **s** no matter what the cause is found by summing over the causes $c$ in Equation 1. Combining that with Equation 3 results in

$$\Pr(\mathbf{s}) = \sum_c \Pr(\mathbf{s}|c)\Pr(c) \ ,$$

$$\Pr(\mathbf{s}) = \sum_c \Pr(c) \prod_s \Pr(s|c)^s (1 - \Pr(s|c))^{1-s} \ . \tag{4}$$

# $\Pr(c|\mathbf{s})$

Substituting the expressions we have just identified (Eqs 3 and 4) into Eq 2 produces a tractable expression for the conditional probability that the cause is $c$ given the observed vector of symptoms $\mathbf{s}$

$$\Pr(c|\mathbf{s}) = \frac{\Pr(c)\prod_s \Pr(s|c)^s(1-\Pr(s|c))^{1-s}}{\sum_c \Pr(c)\prod_s \Pr(s|c)^s(1-\Pr(s|c))^{1-s}} \,. \tag{5}$$

# $\Pr(c|\mathbf{s})$

Substituting the expressions we have just identified (Eqs 3 and 4) into Eq 2 produces a tractable expression for the conditional probability that the cause is $c$ given the observed vector of symptoms $\mathbf{s}$

$$\Pr(c|\mathbf{s}) = \frac{\Pr(c)\prod_s \Pr(s|c)^s(1-\Pr(s|c))^{1-s}}{\sum_c \Pr(c)\prod_s \Pr(s|c)^s(1-\Pr(s|c))^{1-s}}\,. \tag{5}$$

This requires information on both the **presence** and **absence** of a given symptom - i.e. there must be a value for each element in $\mathbf{s}$ and both possible values must have meaning.

# Interpreting the naive Bayes Relationship

$$\Pr(c|\mathbf{s}) = \frac{\Pr(c)\ \prod_s \Pr(s|c)^s(1-\Pr(s|c))^{1-s}}{\sum_c \Pr(c)\ \prod_s \Pr(s|c)^s(1-\Pr(s|c))^{1-s}}$$

$\Pr(c|\mathbf{s})$ is conditional probability of cause of death $c$ given vector of symptoms $\mathbf{s}$

# Interpreting the naive Bayes Relationship

$$\Pr(c|\mathbf{s}) = \frac{\Pr(c) \prod_s \Pr(s|c)^s (1 - \Pr(s|c))^{1-s}}{\sum_c \Pr(c) \prod_s \Pr(s|c)^s (1 - \Pr(s|c))^{1-s}}$$

$\Pr(c|\mathbf{s})$ is conditional probability of cause of death $c$ given vector of symptoms $\mathbf{s}$

$\Pr(c)$ is probability of cause of death $c$ regardless of symptoms – e.g. the **cause-specific mortality fractions**

# Interpreting the naive Bayes Relationship

$$\Pr(c|\mathbf{s}) = \frac{\Pr(c) \prod_s \Pr(s|c)^s (1 - \Pr(s|c))^{1-s}}{\sum_c \Pr(c) \prod_s \Pr(s|c)^s (1 - \Pr(s|c))^{1-s}}$$

$\Pr(c|\mathbf{s})$ is conditional probability of cause of death $c$ given vector of symptoms $\mathbf{s}$

$\Pr(c)$ is probability of cause of death $c$ regardless of symptoms – e.g. the **cause-specific mortality fractions**

$\prod_s \Pr(s|c)^s (1 - \Pr(s|c))^{1-s}$ is probability of vector of symptoms $\mathbf{s}$ given cause of death $c$ *assuming* independence among symptoms

# Interpreting the naive Bayes Relationship

$$\Pr(c|\mathbf{s}) = \frac{\Pr(c) \prod_s \Pr(s|c)^s (1-\Pr(s|c))^{1-s}}{\sum_c \Pr(c) \prod_s \Pr(s|c)^s (1-\Pr(s|c))^{1-s}}$$

$\Pr(c|\mathbf{s})$ is conditional probability of cause of death $c$ given vector of symptoms $\mathbf{s}$

$\Pr(c)$ is probability of cause of death $c$ regardless of symptoms – e.g. the **cause-specific mortality fractions**

$\prod_s \Pr(s|c)^s (1-\Pr(s|c))^{1-s}$ is probability of vector of symptoms $\mathbf{s}$ given cause of death $c$ *assuming* independence among symptoms

$\sum_c \Pr(c) \prod_s \Pr(s|c)^s (1-\Pr(s|c))^{1-s}$ is probability of symptom vector $\mathbf{s}$ regardless of cause of death

# Symptom-Cause Information in Naive Bayes

Recall that ‘Symptom-Cause Information’ **(SCI)** is information that describes the relationship between symptoms and causes

# Symptom-Cause Information in Naive Bayes

Recall that ‘Symptom-Cause Information’ **(SCI)** is information that describes the relationship between symptoms and causes

**SCI come from an external source** such as *physicians* or a *‘gold standard’ dataset* that contains both VA symptoms and a cause(s) of death assigned through an unrelated mechanism, e.g. de-biased physician codes or medical record review

# Symptom-Cause Information in Naive Bayes

Recall that 'Symptom-Cause Information' **(SCI)** is information that describes the relationship between symptoms and causes

**SCI come from an external source** such as *physicians* or a *'gold standard' dataset* that contains both VA symptoms and a cause(s) of death assigned through an unrelated mechanism, e.g. de-biased physician codes or medical record review

SCI appear in the naive Bayes relationship in the form of $\Pr(s|c)$

# Symptom-Cause Information in Naive Bayes

Recall that 'Symptom-Cause Information' **(SCI)** is information that describes the relationship between symptoms and causes

**SCI come from an external source** such as *physicians* or a *'gold standard' dataset* that contains both VA symptoms and a cause(s) of death assigned through an unrelated mechanism, e.g. de-biased physician codes or medical record review

SCI appear in the naive Bayes relationship in the form of $\Pr(s|c)$

$$\Pr(c|\mathbf{s}) = \frac{\Pr(c)\ \prod_s \Pr(s|c)^s(1-\Pr(s|c))^{1-s}}{\sum_c \Pr(c)\ \prod_s \Pr(s|c)^s(1-\Pr(s|c))^{1-s}}$$

# NBC Algorithm for VA

Prabhat Jha's group in Toronto has developed an automated cause-assignment algorithm for VA based on the naive Bayes relationshiop [3] called NBC for verbal autopsy

# NBC Algorithm for VA

Prabhat Jha's group in Toronto has developed an automated cause-assignment algorithm for VA based on the naive Bayes relationshiop [3] called NBC for verbal autopsy

- Uses SCI in the form of $\Pr(s|c)$ derived from Indian death data using Jha group's custom VA tools and validation data
- For each death, algorithm calculates $\Pr(c|\mathbf{s})$ for each cause and identifies the cause with the highest probability as the most likely cause for the death
- Publicly available, open-source R package on CRAN called `https://cran.r-project.org/package=nbc4va`
- Was actually developed and implemented after both InterVA and InSilicoVA

# References I

[1] Peter Byass, Daniel Chandramohan, Samuel J Clark, Lucia D'Ambruoso, Edward Fottrell, Wendy J Graham, Abraham J Herbst, Abraham Hodgson, Sennen Hounton, Kathleen Kahn, et al. Strengthening standardised interpretation of verbal autopsy data: The new interva-4 tool. *Global Health Action*, 5(19281):doi: 10.3402/gha.v5i0.19281, 2012.

[2] Tyler H McCormick, Zehang Richard Li, Clara Calvert, Amelia C Crampin, Kathleen Kahn, and Samuel J Clark. Probabilistic cause-of-death assignment using verbal autopsies. *Journal of the American Statistical Association*, 111(515):1036–1049, 2016.

[3] Pierre Miasnikof, Vasily Giannakeas, Mireille Gomes, Lukasz Aleksandrowicz, Alexander Y Shestopaloff, Dewan Alam, Stephen Tollman, Akram Samarikhalaj, and Prabhat Jha. Naive bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths. *BMC medicine*, 13(1):286, 2015.