Bayes' Rule

Outline

I. Bayes' rule

II. Conditional independence

III. Naïve Bayes model

* Figures are from the textbook site.

I. Bayes' Rule

$$P(a \wedge b) = P(a \mid b)P(b) = P(b \mid a)P(a)$$ (product rule)

I. Bayes' Rule

$$P(a \wedge b) = P(a \mid b)P(b) = P(b \mid a)P(a)$$ (product rule)

$$\Downarrow$$

$$P(a \mid b)P(b) = P(b \mid a)P(a)$$

I. Bayes' Rule

$$P(a \wedge b) = P(a \mid b)P(b) = P(b \mid a)P(a)$$ (product rule)

$$\Downarrow$$

$$P(a \mid b)P(b) = P(b \mid a)P(a)$$

$$\Downarrow$$

$$\boxed{P(b \mid a) = \frac{P(a \mid b)P(b)}{P(a)}}$$ (Bayes' rule)

I. Bayes' Rule

$$P(a \wedge b) = P(a \mid b)P(b) = P(b \mid a)P(a) \qquad \text{(product rule)}$$

$$\Downarrow$$

$$P(a \mid b)P(b) = P(b \mid a)P(a)$$

$$\Downarrow$$

$$\boxed{P(b \mid a) = \frac{P(a \mid b)P(b)}{P(a)}} \qquad \text{(Bayes' rule)}$$

It tells us how often b happens given that a happens,

when we know:

- how often a happens given that b happens, and
- how likely a is on its own, and
- how likely b is on its own.

Bayes' Rule for Multivalued Variables

$$\boldsymbol{P}(Y \mid X) = \frac{\boldsymbol{P}(X \mid Y)\boldsymbol{P}(Y)}{\boldsymbol{P}(X)}$$

The above is a set of equations, each for a pair of possible values of X and Y.

Bayes' Rule for Multivalued Variables

$$\boldsymbol{P}(Y \mid X) = \frac{\boldsymbol{P}(X \mid Y)\boldsymbol{P}(Y)}{\boldsymbol{P}(X)}$$

The above is a set of equations, each for a pair of possible values of X and Y.

A more generalized version conditionalized on some evidence $\boldsymbol{e}$:

$$\boldsymbol{P}(Y \mid X, \boldsymbol{e}) = \frac{\boldsymbol{P}(X \mid Y, \boldsymbol{e})\boldsymbol{P}(Y \mid \boldsymbol{e})}{\boldsymbol{P}(X \mid \boldsymbol{e})}$$

Applying Bayes' Rule

- Perceive as the evidence the *effect* of some unknown *cause*.
- Determine the cause.

$$P\,(\mathit{cause} \mid \mathit{effect}) = \frac{P\,(\mathit{effect} \mid \mathit{cause})P(\mathit{cause})}{P(\mathit{effect})}$$

Applying Bayes' Rule

- Perceive as the evidence the *effect* of some unknown *cause*.
- Determine the cause.

causal relationship

$$P\,(cause \mid effect) = \frac{P\,(effect \mid cause)P(cause)}{P(effect)}$$

Applying Bayes' Rule

- Perceive as the evidence the *effect* of some unknown *cause*.
- Determine the cause.

diagnostic relationship   causal relationship

$$P(\mathit{cause} \mid \mathit{effect}) = \frac{P(\mathit{effect} \mid \mathit{cause})\,P(\mathit{cause})}{P(\mathit{effect})}$$

Applying Bayes' Rule

- Perceive as the evidence the *effect* of some unknown *cause*.
- Determine the cause.

diagnostic relationship　　causal relationship

$$P(\textit{cause} \mid \textit{effect}) = \frac{P(\textit{effect} \mid \textit{cause})P(\textit{cause})}{P(\textit{effect})}$$

Probabilistic information is often available in the form $P(\textit{effect} \mid \textit{cause})$ not $P(\textit{cause} \mid \textit{effect})$.

Applying Bayes' Rule

- Perceive as the evidence the *effect* of some unknown *cause*.
- Determine the cause.

diagnostic relationship causal relationship

$$P(cause \mid effect) = \frac{P(effect \mid cause)P(cause)}{P(effect)}$$

Probabilistic information is often available in the form $P(effect \mid cause)$ not $P(cause \mid effect)$.

The doctor knows $P(symptoms \mid disease)$ and wants to derive a diagnosis $P(disease \mid symptoms)$.

Example

The doctor knows:

- The disease meningitis causes a patient to have a stiff neck 70% of time.
- The prior probability that any patient has meningitis is 1/50,000.
- The prior probability that any patient has a stiff neck is 1/100.

Example

The doctor knows:

- The disease meningitis causes a patient to have a stiff neck 70% of time.
- The prior probability that any patient has meningitis is 1/50,000.
- The prior probability that any patient has a stiff neck is 1/100.

s: The patient has a stiff neck.
m: The patient has meningitis.

Example

The doctor knows:

- The disease meningitis causes a patient to have a stiff neck 70% of time.
- The prior probability that any patient has meningitis is 1/50,000.
- The prior probability that any patient has a stiff neck is 1/100.

s: The patient has a stiff neck.
m: The patient has meningitis.

$$P(s \mid m) = 0.7$$
$$P(m) = 1/50000$$
$$P(s) = 0.01$$

Example

The doctor knows:

- The disease meningitis causes a patient to have a stiff neck 70% of time.
- The prior probability that any patient has meningitis is 1/50,000.
- The prior probability that any patient has a stiff neck is 1/100.

s: The patient has a stiff neck.
m: The patient has meningitis.

$$P(s \mid m) = 0.7$$
$$P(m) = 1/50000$$
$$P(s) = 0.01$$

$$P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)} = \frac{0.7 \cdot 1/50000}{0.01} = 0.0014$$

Example

The doctor knows:

- The disease meningitis causes a patient to have a stiff neck 70% of time.
- The prior probability that any patient has meningitis is 1/50,000.
- The prior probability that any patient has a stiff neck is 1/100.

s: The patient has a stiff neck.
m: The patient has meningitis.

$$P(s \mid m) = 0.7$$
$$P(m) = 1/50000$$
$$P(s) = 0.01$$

$$P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)} = \frac{0.7 \cdot 1/50000}{0.01} = 0.0014$$

Prior Probability of the Evidence

$$P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)}$$

Prior Probability of the Evidence

$$P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)}$$

?

Prior Probability of the Evidence

$$P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)}\ ?$$

We obtain $P(s)$ from computing $P(s \mid m)$ and $P(s \mid \neg m)$:

Prior Probability of the Evidence

$$P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)} \text{ ?}$$

We obtain $P(s)$ from computing $P(s \mid m)$ and $P(s \mid \neg m)$:

$$P(m, s) = P(m \mid s)\, P(s) = P(s \mid m)\, P(m) \quad P(\neg m \mid s)\, P(s) = P(s \mid \neg m)\, P(\neg m)$$

Prior Probability of the Evidence

$$P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)} \text{ ?}$$

We obtain $P(s)$ from computing $P(s \mid m)$ and $P(s \mid \neg m)$:

$$P(m, s) = P(m \mid s)\, P(s) = P(s \mid m)\, P(m) \quad P(\neg m \mid s)\, P(s) = P(s \mid \neg m)\, P(\neg m)$$

$$P(m \mid s)\, P(s) + P(\neg m \mid s)\, P(s) = P(s \mid m)\, P(m) + P(s \mid \neg m)\, P(\neg m)$$

Prior Probability of the Evidence

$$P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)} ?$$

We obtain $P(s)$ from computing $P(s \mid m)$ and $P(s \mid \neg m)$:

$$P(m, s) = P(m \mid s)\, P(s) = P(s \mid m)\, P(m) \quad P(\neg m \mid s)\, P(s) = P(s \mid \neg m)\, P(\neg m)$$

$$\Downarrow$$

$$P(m \mid s)\, P(s) + P(\neg m \mid s)\, P(s) = P(s \mid m)\, P(m) + P(s \mid \neg m)\, P(\neg m)$$

$$\Downarrow$$

$$P\ s = P s \quad m)\ P\ m + P\ s \quad \neg m)\ P\ \neg m$$
$$m\ (s \mid \neg m)\ P(\neg m)$$

Bayes' Rule (General Form)

Apply normalization to Bayes' rule when $P(s)$ is unknown:

$$P(M \mid s) = \alpha \langle P(s \mid m)P(m), P(s \mid \neg m)P(\neg m) \rangle$$

Bayes' Rule (General Form)

Apply normalization to Bayes' rule when $P(s)$ is unknown:

$$P(M \mid s) = \alpha \langle P(s \mid m)P(m), P(s \mid \neg m)P(\neg m) \rangle$$

random variable with
two values m and $\neg m$.

Bayes' Rule (General Form)

Apply normalization to Bayes' rule when $P(s)$ is unknown:

$$P(M \mid s) = \alpha \langle P(s \mid m)P(m), P(s \mid \neg m)P(\neg m) \rangle$$

random variable with two values m and $\neg m$.

$$\alpha = \frac{1}{P(s)} = \frac{1}{P(s \mid m)\, P(m) + P(s \mid \neg m)\, P(\neg m)}$$

Bayes' Rule (General Form)

Apply normalization to Bayes' rule when $P(s)$ is unknown:

$$P(M \mid s) = \alpha \langle P(s \mid m)P(m), P(s \mid \neg m)P(\neg m) \rangle$$

random variable with two values m and $\neg m$.

$$\alpha = \frac{1}{P(s)} = \frac{1}{P(s \mid m)\, P(m) + P(s \mid \neg m)\, P(\neg m)}$$

- $P(\neg m) = 1 - P(m)$

Bayes' Rule (General Form)

Apply normalization to Bayes' rule when $P(s)$ is unknown:

$$P(M \mid s) = \alpha \langle P(s \mid m)P(m), P(s \mid \neg m)P(\neg m) \rangle$$

random variable with two values m and $\neg m$.

$$\alpha = \frac{1}{P(s)} = \frac{1}{P(s \mid m)\, P(m) + P(s \mid \neg m)\, P(\neg m)}$$

- $P(\neg m) = 1 - P(m)$
- $P(s \mid \neg m)$ needs to be estimated instead of $P(s)$.

  Sometimes this is easier, sometimes harder.

Bayes' Rule (General Form)

Apply normalization to Bayes' rule when $P(s)$ is unknown:

$$P(M \mid s) = \alpha \langle P(s \mid m) P(m), P(s \mid \neg m) P(\neg m) \rangle$$

random variable with two values m and $\neg m$.

$$\alpha = \frac{1}{P(s)} = \frac{1}{P(s \mid m)\, P(m) + P(s \mid \neg m)\, P(\neg m)}$$

- $P(\neg m) = 1 - P(m)$
- $P(s \mid \neg m)$ needs to be estimated instead of $P(s)$.

  Sometimes this is easier, sometimes harder.

$$\mathbf{P}(Y \mid X) = \alpha\, \mathbf{P}(X \mid Y) \mathbf{P}(Y)$$

Bayes' Rule (General Form)

Apply normalization to Bayes' rule when $P(s)$ is unknown:

$$P(M \mid s) = \alpha \langle P(s \mid m) P(m), P(s \mid \neg m) P(\neg m) \rangle$$

random variable with two values m and $\neg m$.

$$\alpha = \frac{1}{P(s)} = \frac{1}{P(s \mid m)\, P(m) + P(s \mid \neg m)\, P(\neg m)}$$

- $P(\neg m) = 1 - P(m)$
- $P(s \mid \neg m)$ needs to be estimated instead of $P(s)$.

  Sometimes this is easier, sometimes harder.

$$\boldsymbol{P}(Y \mid X) = \alpha\, \boldsymbol{P}(X \mid Y) \boldsymbol{P}(Y)$$

normalization constant to make the entries in $\boldsymbol{P}(Y \mid X)$ sum to 1.

II. Combining Evidence

What happens when we have two or more pieces of evidences?

- Suppose we know the full joint distribution.

	$toothache$		$\neg toothache$	
	$catch$	$\neg catch$	$catch$	$\neg catch$
$cavity$	0.108	0.012	0.072	0.008
$\neg cavity$	0.016	0.064	0.144	0.576

II. Combining Evidence

What happens when we have two or more pieces of evidences?

- Suppose we know the full joint distribution.

	$toothache$		$\neg toothache$	
	$catch$	$\neg catch$	$catch$	$\neg catch$
$cavity$	0.108	0.012	0.072	0.008
$\neg cavity$	0.016	0.064	0.144	0.576

$$\mathbf{P}(Cavity \mid toothache \wedge catch) = \alpha\langle 0.108, 0.016\rangle \approx \langle 0.871, 0.129\rangle$$

II. Combining Evidence

What happens when we have two or more pieces of evidences?

- Suppose we know the full joint distribution.

	$toothache$		$\neg toothache$	
	$catch$	$\neg catch$	$catch$	$\neg catch$
$cavity$	0.108	0.012	0.072	0.008
$\neg cavity$	0.016	0.064	0.144	0.576

$$\boldsymbol{P}(Cavity \mid toothache \wedge catch) = \alpha\langle 0.108, 0.016\rangle \approx \langle 0.871, 0.129\rangle$$

♠ Does not scale up to a large number of evidence variables.

II. Combining Evidence

What happens when we have two or more pieces of evidences?

- Suppose we know the full joint distribution.

	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

$$\mathbf{P}(Cavity \mid toothache \wedge catch) = \alpha\langle 0.108, 0.016\rangle \approx \langle 0.871, 0.129\rangle$$

♠ Does not scale up to a large number of evidence variables.

- Apply Bayes' rule:

$$\mathbf{P}(Cavity \mid toothache \wedge catch) = \alpha \mathbf{P}(toothache \wedge catch \mid Cavity)\, \mathbf{P}(Cavity)$$

II. Combining Evidence

What happens when we have two or more pieces of evidences?

- Suppose we know the full joint distribution.

	toothache		$\neg$*toothache*	
	catch	$\neg$*catch*	*catch*	$\neg$*catch*
cavity	0.108	0.012	0.072	0.008
$\neg$*cavity*	0.016	0.064	0.144	0.576

$$\mathbf{P}(Cavity \mid toothache \wedge catch) = \alpha\langle 0.108, 0.016\rangle \approx \langle 0.871, 0.129\rangle$$

♠ Does not scale up to a large number of evidence variables.

- Apply Bayes' rule:

$$\mathbf{P}(Cavity \mid toothache \wedge catch) = \alpha \mathbf{P}(toothache \wedge catch \mid Cavity)\, \mathbf{P}(Cavity)$$

n (Boolean) evidence variables: *Toothache*, *Catch*, *X-rays*, *Diet*, …

II. Combining Evidence

What happens when we have two or more pieces of evidences?

- Suppose we know the full joint distribution.

	toothache		$\neg$*toothache*	
	catch	$\neg$*catch*	*catch*	$\neg$*catch*
cavity	0.108	0.012	0.072	0.008
$\neg$*cavity*	0.016	0.064	0.144	0.576

$$\mathbf{P}(Cavity \mid toothache \wedge catch) = \alpha\langle 0.108, 0.016\rangle \approx \langle 0.871, 0.129\rangle$$

♠ Does not scale up to a large number of evidence variables.

- Apply Bayes' rule:

$$\mathbf{P}(Cavity \mid toothache \wedge catch) = \alpha \mathbf{P}(toothache \wedge catch \mid Cavity)\, \mathbf{P}(Cavity)$$

n (Boolean) evidence variables: *Toothache*, *Catch*, *X-rays*, *Diet*, …

$\Longrightarrow$ 2^n possible combinations of observed values needed to determine $\mathbf{P}(toothache \wedge catch \wedge \cdots \mid Cavity)$!

Conditional Independence

- *Toothache* and *Catch* are independent *given the presence or absence of a cavity*.

Conditional Independence

- *Toothache* and *Catch* are independent *given the presence or absence of a cavity*.
  - Each is directly caused by the cavity.
  - Neither has a direct effect on the other.

    Toothache depends on the state of the nerves in the tooth.

    The probe's accuracy depends on the dentist's skill irrelevant to the toothache.

Conditional Independence

- *Toothache* and *Catch* are independent *given the presence or absence of a cavity*.
  - Each is directly caused by the cavity.
  - Neither has a direct effect on the other.

    Toothache depends on the state of the nerves in the tooth.

    The probe's accuracy depends on the dentist's skill irrelevant to the toothache.

- Conditional independence of *toothache* and *catch* given *Cavity*:

$$\boldsymbol{P}(toothache \wedge catch \mid Cavity) = \boldsymbol{P}(toothache \mid Cavity)\, \boldsymbol{P}(catch \mid Cavity)$$

Conditional Independence

- *Toothache* and *Catch* are independent *given the presence or absence of a cavity*.
  - Each is directly caused by the cavity.
  - Neither has a direct effect on the other.

    Toothache depends on the state of the nerves in the tooth.

    The probe's accuracy depends on the dentist's skill irrelevant to the toothache.

- Conditional independence of *toothache* and *catch* given *Cavity*:

$$\boldsymbol{P}(toothache \wedge catch \mid Cavity) = \boldsymbol{P}(toothache \mid Cavity)\, \boldsymbol{P}(catch \mid Cavity)$$

Bayes' rule: $\boldsymbol{P}(Cavity \mid toothache \wedge catch) = \alpha\, \boldsymbol{P}(toothache \wedge catch \mid Cavity)\, \boldsymbol{P}(Cavity)$

Conditional Independence

- *Toothache* and *Catch* are independent *given the presence or absence of a cavity*.
  - Each is directly caused by the cavity.
  - Neither has a direct effect on the other.

    Toothache depends on the state of the nerves in the tooth.

    The probe's accuracy depends on the dentist's skill irrelevant to the toothache.

- Conditional independence of *toothache* and *catch* given *Cavity*:

$$\boldsymbol{P}(toothache \wedge catch \mid Cavity) = \boldsymbol{P}(toothache \mid Cavity)\, \boldsymbol{P}(catch \mid Cavity)$$

Bayes' rule: $\boldsymbol{P}(Cavity \mid toothache \wedge catch) = \alpha \boxed{\boldsymbol{P}(toothache \wedge catch \mid Cavity)}\, \boldsymbol{P}(Cavity)$

Conditional Independence

- *Toothache* and *Catch* are independent *given the presence or absence of a cavity*.
  - ♦ Each is directly caused by the cavity.
  - ♦ Neither has a direct effect on the other.

    Toothache depends on the state of the nerves in the tooth.

    The probe's accuracy depends on the dentist's skill irrelevant to the toothache.

- Conditional independence of *toothache* and *catch* given *Cavity*:

$$\mathbf{P}(toothache \wedge catch \mid Cavity) = \mathbf{P}(toothache \mid Cavity)\, \mathbf{P}(catch \mid Cavity)$$

Bayes' rule: $\mathbf{P}(Cavity \mid toothache \wedge catch) = \alpha\, \boxed{\mathbf{P}(toothache \wedge catch \mid Cavity)}\, \mathbf{P}(Cavity)$

$$\Downarrow$$

$$\mathbf{P}(Cavity \mid toothache \wedge catch)$$
$$= \alpha \mathbf{P}(toothache \mid Cavity)\, \mathbf{P}(catch \mid Cavity) \mathbf{P}(Cavity)$$

Conditional Independence

- *Toothache* and *Catch* are independent *given the presence or absence of a cavity*.
  - Each is directly caused by the cavity.
  - Neither has a direct effect on the other.

    Toothache depends on the state of the nerves in the tooth.

    The probe's accuracy depends on the dentist's skill irrelevant to the toothache.

- Conditional independence of *toothache* and *catch* given *Cavity*:

$$\mathbf{P}(toothache \wedge catch \mid Cavity) = \mathbf{P}(toothache \mid Cavity)\ \mathbf{P}(catch \mid Cavity)$$

Bayes' rule: $\mathbf{P}(Cavity \mid toothache \wedge catch) = \alpha \boxed{\mathbf{P}(toothache \wedge catch \mid Cavity)} \mathbf{P}(Cavity)$

$$\Downarrow$$

$$\mathbf{P}(Cavity \mid toothache \wedge catch)$$
$$= \alpha \mathbf{P}(toothache \mid Cavity)\ \mathbf{P}(catch \mid Cavity) \mathbf{P}(Cavity)$$

Definition of Conditional Independence

Independence of two variables X and Y given a third variable Z:

$$P(X, Y \mid Z) = P(X \mid Z)\, P(Y \mid Z)$$

Definition of Conditional Independence

Independence of two variables X and Y given a third variable Z:

$$P(X, Y \mid Z) = P(X \mid Z)\, P(Y \mid Z)$$

$\boldsymbol{P}(\mathit{Toothache}, \mathit{Catch} \mid \mathit{Cavity}) = \boldsymbol{P}(\mathit{Toothache} \mid \mathit{Cavity})\, \boldsymbol{P}(\mathit{Catch} \mid \mathit{Cavity})$

// asserts independence for all possible combinations of values of
// *Toothache* and *Catch*.

Definition of Conditional Independence

Independence of two variables X and Y given a third variable Z:

$$P(X, Y \mid Z) = P(X \mid Z)\, P(Y \mid Z)$$

$\boldsymbol{P}(\mathit{Toothache}, \mathit{Catch} \mid \mathit{Cavity}) = \boldsymbol{P}(\mathit{Toothache} \mid \mathit{Cavity})\, \boldsymbol{P}(\mathit{Catch} \mid \mathit{Cavity})$

// asserts independence for all possible combinations of values of
// *Toothache* and *Catch*.

Equivalent forms:

Definition of Conditional Independence

Independence of two variables X and Y given a third variable Z:

$$P(X, Y \mid Z) = P(X \mid Z)\, P(Y \mid Z)$$

$\boldsymbol{P}(\mathit{Toothache}, \mathit{Catch} \mid \mathit{Cavity}) = \boldsymbol{P}(\mathit{Toothache} \mid \mathit{Cavity})\, \boldsymbol{P}(\mathit{Catch} \mid \mathit{Cavity})$

// asserts independence for all possible combinations of values of
// *Toothache* and *Catch*.

Equivalent forms:

$P(X \mid Y, Z) = P(X \mid Z)$ and $P(Y \mid X, Z) = P(Y \mid Z)$

Definition of Conditional Independence

Independence of two variables X and Y given a third variable Z:

$$P(X, Y \mid Z) = P(X \mid Z)\, P(Y \mid Z)$$

$\boldsymbol{P}(\mathit{Toothache}, \mathit{Catch} \mid \mathit{Cavity}) = \boldsymbol{P}(\mathit{Toothache} \mid \mathit{Cavity})\, \boldsymbol{P}(\mathit{Catch} \mid \mathit{Cavity})$

// asserts independence for all possible combinations of values of
// *Toothache* and *Catch*.

Equivalent forms:

$$P(X \mid Y, Z) = P(X \mid Z) \quad \text{and} \quad P(Y \mid X, Z) = P(Y \mid Z)$$

Decomposition into smaller conditional assertions.

Decomposition of Joint Distribution

$$\mathbf{P}(\mathit{Toothache}, \mathit{Catch}, \mathit{Cavity})$$

Decomposition of Joint Distribution

$$\mathbf{P}(Toothache, Catch, Cavity)$$

$$= \mathbf{P}(Toothache, Catch \mid Cavity)\ \mathbf{P}(Cavity)$$

Decomposition of Joint Distribution

$$\boldsymbol{P}(Toothache, Catch, Cavity)$$

$$= \boldsymbol{P}(Toothache, Catch \mid Cavity)\ \boldsymbol{P}(Cavity)$$

$$= \boldsymbol{P}(Toothache\ \ aiy\ \boldsymbol{P}(ach\ \ aiy\ \boldsymbol{P}(aiy)$$

$$\text{ty ty}\ \ \boldsymbol{P}(Catch \mid Cavity)\ \boldsymbol{P}(Cavity)$$

Decomposition of Joint Distribution

$$\boldsymbol{P}(Toothache, Catch, Cavity)$$

$$= \boldsymbol{P}(Toothache, Catch \mid Cavity)\ \boldsymbol{P}(Cavity)$$

$$= \boldsymbol{P}(Toothache\ \ aiy\ \boldsymbol{P}(ach\ \ aiy\ \boldsymbol{P}(aiy)$$

$$\underbrace{\text{ty ty}\ \ \boldsymbol{P}(Catch \mid Cavity)\ \boldsymbol{P}(Cavity)}$$

3 tables of dimensions $2 \times 2, 2 \times 2$, and 2×1
with a total of $2 + 2 + 1 = 5$ independent numbers
(which appear in the first row of every table).

Decomposition of Joint Distribution

$$\boldsymbol{P}(Toothache, Catch,\ Cavity)$$

$$= \boldsymbol{P}(Toothache, Catch \mid Cavity)\ \boldsymbol{P}(Cavity)$$

$$= \boldsymbol{P}(Toothache\ \ aiy\ \ \boldsymbol{P}(ach\ \ aiy\ \ \boldsymbol{P}(aiy)$$

$$\underbrace{\text{ty ty}\ \ \boldsymbol{P}(Catch \mid Cavity)\ \boldsymbol{P}(Cavity)}$$

3 tables of dimensions $2 \times 2, 2 \times 2$, and 2×1
with a total of $2 + 2 + 1 = 5$ independent numbers
(which appear in the first row of every table).

With n symptoms, the size grows as $O(n)$ instead of $O(2^n)$.

Decomposition of Joint Distribution

$$\mathbf{P}(Toothache, Catch, Cavity)$$

$$= \mathbf{P}(Toothache, Catch \mid Cavity)\ \mathbf{P}(Cavity)$$

$$= \underbrace{\mathbf{P}(Toothache \mid Cavity)\ \mathbf{P}(Catch \mid Cavity)\ \mathbf{P}(Cavity)}$$

3 tables of dimensions 2×2, 2×2, and 2×1
with a total of $2 + 2 + 1 = 5$ independent numbers
(which appear in the first row of every table).

With n symptoms, the size grows as $O(n)$ instead of $O(2^n)$.

- Conditional independence assertions allow probabilistic systems to scale up.
- They are more commonly available than absolute independence assertions.
- Decomposition of large probabilistic domains through conditional independence is one of the most important recent developments in AI.

III. Naive Bayes Model

$$\boldsymbol{P}(Cause, Effect_1, \dots, Effect_n) = \boldsymbol{P}(Cause)\prod_i \boldsymbol{P}(Effect_i \mid Cause)$$

III. Naive Bayes Model

$$\boldsymbol{P}(Cause, Effect_1, \ldots, Effect_n) = \boldsymbol{P}(Cause) \prod_i \boldsymbol{P}(Effect_i \mid Cause)$$

- often works well in cases where the "effect" variables $Effect_1, \ldots, Effect_n$ are not strictly independent.

III. Naive Bayes Model

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause) \prod_i \mathbf{P}(Effect_i \mid Cause)$$

- often works well in cases where the “effect” variables $Effect_1, \ldots, Effect_n$ are not strictly independent.

$$\mathbf{P}(Cause \mid \mathbf{e}) = \alpha \sum_{\mathbf{y}} \mathbf{P}(Cause, \mathbf{e}, \mathbf{y})$$

III. Naive Bayes Model

$$\boldsymbol{P}(Cause, Effect_1, \ldots, Effect_n) = \boldsymbol{P}(Cause) \prod_i \boldsymbol{P}(Effect_i \mid Cause)$$

- often works well in cases where the "effect" variables $Effect_1, \ldots, Effect_n$ are not strictly independent.

$$\begin{aligned}\boldsymbol{P}(Cause \mid \boldsymbol{e}) &= \alpha \sum_{\boldsymbol{y}} \boldsymbol{P}(Cause, \boldsymbol{e}, \boldsymbol{y}) \\ &= \alpha \sum_{\boldsymbol{y}} \boldsymbol{P}(Cause) \boldsymbol{P}(\boldsymbol{y} \mid Cause) \left(\prod_j \boldsymbol{P}(e_j \mid Cause) \right)\end{aligned}$$

III. Naive Bayes Model

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause) \prod_i \mathbf{P}(Effect_i \mid Cause)$$

- often works well in cases where the "effect" variables $Effect_1, \ldots, Effect_n$ are not strictly independent.

$$\begin{aligned}\mathbf{P}(Cause \mid \mathbf{e}) &= \alpha \sum_{\mathbf{y}} \mathbf{P}(Cause, \mathbf{e}, \mathbf{y}) \\ &= \alpha \sum_{\mathbf{y}} \mathbf{P}(Cause)\mathbf{P}(\mathbf{y} \mid Cause) \left(\prod_j \mathbf{P}(e_j \mid Cause) \right) \\ &= \alpha \mathbf{P}(Cause) \left(\prod_j \mathbf{P}(e_j \mid Cause) \right) \sum_{\mathbf{y}} \mathbf{P}(\mathbf{y} \mid Cause)\end{aligned}$$

III. Naive Bayes Model

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause) \prod_i \mathbf{P}(Effect_i \mid Cause)$$

- often works well in cases where the "effect" variables $Effect_1, \ldots, Effect_n$ are not strictly independent.

$$\begin{aligned}\mathbf{P}(Cause \mid \boldsymbol{e}) &= \alpha \sum_{\boldsymbol{y}} \mathbf{P}(Cause, \boldsymbol{e}, \boldsymbol{y}) \\ &= \alpha \sum_{\boldsymbol{y}} \mathbf{P}(Cause)\mathbf{P}(\boldsymbol{y} \mid Cause) \left(\prod_j \mathbf{P}(e_j \mid Cause)\right) \\ &= \alpha \mathbf{P}(Cause) \left(\prod_j \mathbf{P}(e_j \mid Cause)\right) \underbrace{\sum_{\boldsymbol{y}} \mathbf{P}(\boldsymbol{y} \mid Cause)}_{=1}\end{aligned}$$

III. Naive Bayes Model

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause) \prod_i \mathbf{P}(Effect_i \mid Cause)$$

- often works well in cases where the "effect" variables $Effect_1, \ldots, Effect_n$ are not strictly independent.

$$\begin{aligned}
\mathbf{P}(Cause \mid \mathbf{e}) &= \alpha \sum_{\mathbf{y}} \mathbf{P}(Cause, \mathbf{e}, \mathbf{y}) \\
&= \alpha \sum_{\mathbf{y}} \mathbf{P}(Cause)\mathbf{P}(\mathbf{y} \mid Cause) \left(\prod_j \mathbf{P}(e_j \mid Cause) \right) \\
&= \alpha \mathbf{P}(Cause) \left(\prod_j \mathbf{P}(e_j \mid Cause) \right) \underbrace{\sum_{\mathbf{y}} \mathbf{P}(\mathbf{y} \mid Cause)}_{=1} \\
&= \alpha \mathbf{P}(Cause) \left(\prod_j \mathbf{P}(e_j \mid Cause) \right)
\end{aligned}$$

(cont'd)

$$\boldsymbol{P}(Cause \mid \boldsymbol{e}) = \alpha \boldsymbol{P}(Cause)\left(\prod_j \boldsymbol{P}(e_j \mid Cause)\right)$$

Calculate the probability distribution of the causes from observed effects:

- Take each possible cause.
- Multiply its prior probability by the product of the conditional probabilities of the observed effects given that cause.
- Normalize the result.

(cont'd)

$$\mathbf{P}(Cause \mid \mathbf{e}) = \alpha \mathbf{P}(Cause) \left(\prod_j \mathbf{P}(e_j \mid Cause) \right)$$

Calculate the probability distribution of the causes from observed effects:

- Take each possible cause.
- Multiply its prior probability by the product of the conditional probabilities of the observed effects given that cause.
- Normalize the result.

♦ *Linear run time* in the number of observed effects only.

(cont'd)

$$\mathbf{P}(Cause \mid \mathbf{e}) = \alpha \mathbf{P}(Cause) \left(\prod_j \mathbf{P}(e_j \mid Cause) \right)$$

Calculate the probability distribution of the causes from observed effects:

- Take each possible cause.
- Multiply its prior probability by the product of the conditional probabilities of the observed effects given that cause.
- Normalize the result.

♦ *Linear run time* in the number of observed effects only.

♦ The number of unobserved effects is irrelevant no matter how large it is (as in medicine).

Text Classification

Problem Given a text, decide which of a predefined set of classes or categories it belongs to.

Text Classification

Problem Given a text, decide which of a predefined set of classes or categories it belongs to.

“cause” → *Category* with range, e.g.,

{*news, sports, business, weather, entertainment* }

Text Classification

Problem Given a text, decide which of a predefined set of classes or categories it belongs to.

"cause" → *Category* with range, e.g.,

{*news, sports, business, weather, entertainment* }

"effects" → $\mathit{HasWord}_i$: presence or absence of the i^{th} keyword.

Text Classification

Problem Given a text, decide which of a predefined set of classes or categories it belongs to.

"cause" → *Category* with range, e.g.,

{*news, sports, business, weather, entertainment* }

"effects" → $HasWord_i$: presence or absence of the i^{th} keyword.

Example sentences:

Text Classification

Problem Given a text, decide which of a predefined set of classes or categories it belongs to.

"cause" → *Category* with range, e.g.,

{*news, sports, business, weather, entertainment* }

"effects" → *HasWord*$_i$: presence or absence of the i^{th} keyword.

Example sentences:

1. Stocks rallied on Monday, with major indexes gaining 1% as optimism persisted over the first quarter earnings season.
2. Heavy rain continued to pound much of the east coast on Monday, with flood warnings issued in New York City and other locations..

Text Classification

Problem Given a text, decide which of a predefined set of classes or categories it belongs to.

"cause" → *Category* with range, e.g.,

{*news, sports, business, weather, entertainment* }

"effects" → $HasWord_i$: presence or absence of the i^{th} keyword.

Example sentences:

1. Stocks rallied on Monday, with major indexes gaining 1% as optimism persisted over the first quarter earnings season.
2. Heavy rain continued to pound much of the east coast on Monday, with flood warnings issued in New York City and other locations..

Classify each sentence into a Category.

Classification (cont'd)

- Prior probabilities: $\boldsymbol{P}(Category)$
- Conditional probabilities: $\boldsymbol{P}(HasWord_i \mid Category)$

Classification (cont'd)

- Prior probabilities: $\boldsymbol{P}(Category)$
- Conditional probabilities: $\boldsymbol{P}(HasWord_i \mid Category)$
  - ♣ $\boldsymbol{P}(Category = c) \approx$ fraction of all previously seen documents that are of c.

Classification (cont'd)

- Prior probabilities: $\boldsymbol{P}(Category)$
- Conditional probabilities: $\boldsymbol{P}(HasWord_i \mid Category)$

  ♣ $\boldsymbol{P}(Category = c) \approx$ fraction of all previously seen documents that are of c.

  $\boldsymbol{P}(Category = weather) = 0.09$ // 9% of articles are about weather.

Classification (cont'd)

- Prior probabilities: $\boldsymbol{P}(\textit{Category})$
- Conditional probabilities: $\boldsymbol{P}(\textit{HasWord}_i \mid \textit{Category})$

♣ $\boldsymbol{P}(\textit{Category} = c) \approx$ fraction of all previously seen documents that are of c.

$\boldsymbol{P}(\textit{Category} = \textit{weather}) = 0.09$ // 9% of articles are about weather.

♣ $\boldsymbol{P}(\textit{HasWord}_i \mid \textit{Category}) \approx$ fraction of documents of each category that contain word i.

Classification (cont'd)

- Prior probabilities: $\boldsymbol{P}(\textit{Category})$
- Conditional probabilities: $\boldsymbol{P}(\textit{HasWord}_i \mid \textit{Category})$

♣ $\boldsymbol{P}(\textit{Category} = c) \approx$ fraction of all previously seen documents that are of c.

$\boldsymbol{P}(\textit{Category} = \textit{weather}) = 0.09$ // 9% of articles are about weather.

♣ $\boldsymbol{P}(\textit{HasWord}_i \mid \textit{Category}) \approx$ fraction of documents of each category that contain word i.

$P(\textit{HasWord}_6 = \textit{true} \mid \textit{Category} = \textit{business}) \approx 0.37$
// 37% of articles about business contain word 6, “stocks”.

Classification Algorithm

- Check which key words (i.e., effects) appear.

Classification Algorithm

- Check which key words (i.e., effects) appear.
- Compute their posterior probability distribution over categories (i.e. causes).

Classification Algorithm

- Check which key words (i.e., effects) appear.
- Compute their posterior probability distribution over categories (i.e. causes).

$$\begin{aligned}&\boldsymbol{P}(Category \mid HasWords)\\&= \alpha \boldsymbol{P}(Category)\left(\prod_j \boldsymbol{P}(HasWord_j \mid Category)\right)\end{aligned}$$

Classification Algorithm

- Check which key words (i.e., effects) appear.
- Compute their posterior probability distribution over categories (i.e. causes).

$$\begin{aligned}&\mathbf{P}(Category \mid HasWords) \\ &= \alpha \mathbf{P}(Category)\left(\prod_j \mathbf{P}(HasWord_j \mid Category)\right)\end{aligned}$$

$$e.g.,\ HasWords = \langle HasWord_1 = true, HasWord_2 = false, \ldots \rangle$$

Classification Algorithm

- Check which key words (i.e., effects) appear.
- Compute their posterior probability distribution over categories (i.e. causes).

$$\begin{aligned} &\mathbf{P}(Category \mid HasWords) \\ &= \alpha \mathbf{P}(Category) \left(\prod_j \mathbf{P}(HasWord_j \mid Category) \right) \end{aligned}$$

$$e.g., HasWords = \langle \underbrace{HasWord_1 = true, HasWord_2 = false, \ldots}_{\text{appearances/disappearances of the key words.}} \rangle$$

Other Applications of Naïve Bayes Models

- Language determination (to detect the language a text is written in)
- Spam filtering (to identify spam e-mails)
- Sentiment analysis (to identify positive and negative customer sentiments in social media)
- Real-time prediction (because they are very fast)
- Recommendation systems (to filter unseen information and predict whether a user would like a given resource or not)

Other Applications of Naïve Bayes Models

- ♦ Language determination (to detect the language a text is written in)
- ♦ Spam filtering (to identify spam e-mails)
- ♦ Sentiment analysis (to identify positive and negative customer sentiments in social media)
- ♦ Real-time prediction (because they are very fast)
- ♦ Recommendation systems (to filter unseen information and predict whether a user would like a given resource or not)

Naïve Bayes models are not used in

- ♠ Medical diagnosis (which requires more sophisticated models)