## Last Time

- Introduction: Probability, Statistics, Variance
  - Deming’s Experimental Procedure
  - Anomaly Detection: Western Electronic Rules
  - Mechanistic vs. Empirical Models
- Terminology
  - Random Experiments, Sample Spaces, Events
  - Interpretations and Axioms of Probability
  - Addition Rules
  - Conditional Probability
  - Multiplication and Total Probability Rules



## Last Time

- Experiment
  - Random experiment: Throwing a dice, coin toss etc.
- Outcome: Result of an experiment
  - Dice:{1}, Coin:{Head}
- Sample Space: Set of all outcomes
  - Dice Sample Space: {1,2,3,4,5,6}
- Event: Set of outcomes. Events are subsets of the sample space.
  - Dice: Getting an even number : {2,4,6}
- Probability: is the likelihood or chance that a particular ***outcome*** or ***event*** from a random experiment will occur.
  - Dice: P(Getting an event number) = 3/6

## Last Time Ctd.

If $S$ is the sample space and $E$ is any event in the random experiment,

- $P(S) = 1$
- $0 \leq P(E) \leq 1$
- $P(\varnothing) = 0$ and $P(E') = 1 - P(E)$
- If $E_1$ is contained in $E_2$, then $P(E_1) \leq P(E_2)$
- For any two events $E_1$ and $E_2$
  - $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E)$
- If $E_1$ and $E_2$ are mutually exclusive (i.e. independent), then
  - $P(E_1 \cap E) = 0$ and $P(E_1 \cup E_2) = P(E_1) + P(E_2)$

3

## Last Time Ctd.

- Conditional Probability: $P(B \mid A)$ is the probability of event $B$ occurring, given that event $A$ has already occurred.
- The conditional probability of an event $B$ given an event $A$, denoted as $P(B \mid A)$, is:

  $P(B \mid A) = P(A \cap B) / P(A)$ for $P(A) > 0$.
- The conditional probability can be rewritten to generalize a multiplication rule.

$$P(A \cap B) = P(B|A) \cdot P(A) = P(A|B) \cdot P(B)$$

4

# Today

- Independence
- Bayes Theorem
- Classification
  - Naive Bayes as a Classifier

5

# Example: Sampling Without Enumeration

- A data set contains 50 transactions 10 made by Customer 1 and 40 made by Customer 2. If 2 transactions are selected randomly*,

  a) What is the probability that the 2nd transaction came from Customer 2, given that the 1st transaction came from Customer 1?

  - $P(E_1)= P(1^{st}$ trans. came from C1) = 10/50
  - $P(E_2 \mid E_1) = P(2^{nd}$ trans. came from C2 given that 1st came from C1) = 40/49

  b) What is the probability that the 1st trans. came from C1 and the 2nd trans came from C2?

  $P(E_1 \cap E_2) = P(1^{st}$ part came from C1 and 2nd came from C2)
  $= (10/50)\cdot(40/49) = 8/49$

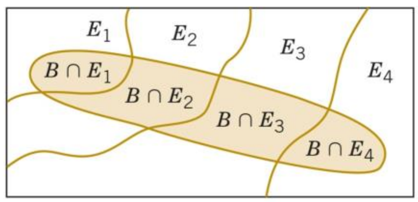
*Selected randomly implies that at each step of the sample, the items remain in the batch are equally likely to be selected.

6

# Total Probability Rule (Multiple Events)

- A collection of sets $E_1, E_2, \ldots E_k$ such that $E_1 \cup E_2 \cup \ldots\ldots \cup E_k = S$ is said to be **exhaustive**.
- Assume $E_1, E_2, \ldots E_k$ are $k$ mutually exclusive and exhaustive. Then

$$P(B) = P(B \cap E_1) + P(B \cap E_2) + \ldots + P(B \cap E_k)$$
$$= P(B \mid E_1) \cdot P(E_1) + P(B \mid E_2) \cdot P(E_2) + \ldots + P(B \mid E_k) \cdot P(E_k)$$




$$B = (B \cap E_1) \cup (B \cap E_2) \cup (B \cap E_3) \cup (B \cap E_4)$$



# Example: Segment based churn

Continuing the discussion of churn, find the probability of churn of the population.

| Probability of Churn | Segment | Probability of Segment |
|---|---|---|
| 0,100 | Age<25 | 0,2 |
| 0,010 | 25<Age<35 | 0,3 |
| 0,001 | Age>35 | 0,5 |

# Event Independence

- Two events are independent ***iff*** any one of the following equivalent statements is true:
  1. $P(A \mid B) = P(A)$
  2. $P(B \mid A) = P(B)$
  3. $P(A \cap B) = P(A) \cdot P(B)$
- This means that occurrence of one event has no impact on the probability of occurrence of the other event.



# Example: Independent or not?

Table 1 provides an example of 400 phones classified by battery flaws and as (functionally) defective. Suppose that Table 2 represents a different situation (i.e. screen flaws vs. functional problem). Let *F* denote the event that the phone has flaws. Let *D* denote the event that the phone is defective. Check independency of F & D for the cases presented in Table 1 and Table 2

| TABLE 1 Parts Classified | | | | TABLE 2 Parts Classified (data chg'd) | | | |
|---|---|---|---|---|---|---|---|
| | Battery Flaws | | | | Screen Flaws | | |
| Defective | Yes ($F$) | No ($F'$) | Total | Defective | Yes ($F$) | No ($F'$) | Total |
| Yes ($D$) | 10 | 18 | 28 | Yes ($D$) | 2 | 18 | 20 |
| No ($D'$) | 30 | 342 | 372 | No ($D'$) | 38 | 342 | 380 |
| Total | 40 | 360 | 400 | Total | 40 | 360 | 400 |

# Example: Independent or not?

Table 1 provides an example of 400 phones classified by battery flaws and as (functionally) defective. Suppose that Table 2 represents a different situation (i.e. screen flaws vs. functional problem). Let $F$ denote the event that the phone has flaws. Let $D$ denote the event that the phone is defective. Check independency of F & D for the cases presented in Table 1 and Table 2

| TABLE 1 Parts Classified | | | | TABLE 2 Parts Classified (data chg'd) | | | |
|---|---|---|---|---|---|---|---|
| | Battery Flaws | | | | Screen Flaws | | |
| Defective | Yes ($F$) | No ($F'$) | Total | Defective | Yes ($F$) | No ($F'$) | Total |
| Yes ($D$) | 10 | 18 | 28 | Yes ($D$) | 2 | 18 | 20 |
| No ($D'$) | 30 | 342 | 372 | No ($D'$) | 38 | 342 | 380 |
| Total | 40 | 360 | 400 | Total | 40 | 360 | 400 |
| | | | | | | | |
| | $P(D\|F)$ = | 10/40 = | 0,25 | | $P(D\|F)$ = | 2/40 = | 0,05 |
| | $P(D)$ = | 28/400 = | 0,10 | | $P(D)$ = | 20/400 = | 0,05 |
| | | | not same | | | | same |
| | Events $D$ & $F$ are **dependent** | | | | Events $D$ & $F$ are **independent** | | |



# Independence with Multiple Events

The events $E_1$, $E_2$, … , $E_k$ are independent if and only if (*iff*), for any subset of these events:

$$P(E_{i1} \cap E_{i2} \cap \dots , \cap E_{ik}) = P(E_{i1}) \cdot P(E_{i2}) \cdot \dots \cdot P(E_{ik})$$

Note the *iff* above. The proposition is valid for both directions:

Indep. implies $P(E_{i1} \cap E_{i2} \cap \dots , \cap E_{ik}) = P(E_{i1}) \cdot P(E_{i2}) \cdot \dots \cdot P(E_{ik})$

$P(E_{i1} \cap E_{i2} \cap \dots , \cap E_{ik}) = P(E_{i1}) \cdot P(E_{i2}) \cdot \dots \cdot P(E_{ik})$ implies indep.

# Example: Probability of Churn

Assume the churn probability of a customer is 0.01 and that the customers are independent; that is, the probability that a customer churns does not depend on the behavior of any of the other customers. If 15 customers are analyzed, what is the probability that no customer has churned?

Let $E_i$ denote the event that the $i^{th}$ customer has churned, i = 1, 2, …,15.

Then , $P(E_i) = 0.99$.

The required probability is $P(E_1 \cap E_2 \cap \ldots \cap E_{15})$.

From the assumption of independence,

$$P(E_1 \cap E_2 \cap \ldots \cap E_{15}) = P(E_1) \cdot P(E_2) \cdot \ldots \cdot P(E_{15})$$
$$= (0.99)^{15}$$
$$= 0.86.$$

13

# Bayes' Theorem

- Thomas Bayes (1702-1761) was an English mathematician.
- His idea was that we observe conditional probabilities through prior information.
- Bayes' theorem (Bayes' rule) describes the probability of an event based on conditions that might be related to the event.
  - Suppose we want to know patient A's risk of having hearth attack. Assume the patient is 50 years old. If hearth attack risk is related to age, information about patient's age can be used to assess chance of having hearth attack using Bayes' Theorem.
  - Check the following notation, where $A$: hearth attack, $B$: age
- Bayes' theoren $P(A \mid B) = \frac{P(B \mid A) \bullet P(A)}{P(B)}$ for $P(B) > 0$

14

# Example

The conditional probability that a high level of contamination was present when a failure occurred is to be determined.

The information from previous example is summarized here.

| Probability of Failure | Level of Contamination | Probability of Level |
|---|---|---|
| 0.1 | High | 0.2 |
| 0.005 | Not High | 0.8 |

Let *F* denote the event that the product fails, and let *H* denote the event that the chip is exposed to high levels of contamination. The requested probability is *P*(H|*F*).

$$P(H \mid F) = \frac{P(F \mid H) \cdot P(H)}{P(F)} = \frac{0.10 \cdot 0.20}{0.024} = 0.83$$

$$P(F) = P(F \mid H) \cdot P(H) + P(F \mid H') \cdot P(H')$$
$$= 0.1 \cdot 0.2 + 0.005 \cdot 0.8 = 0.024$$



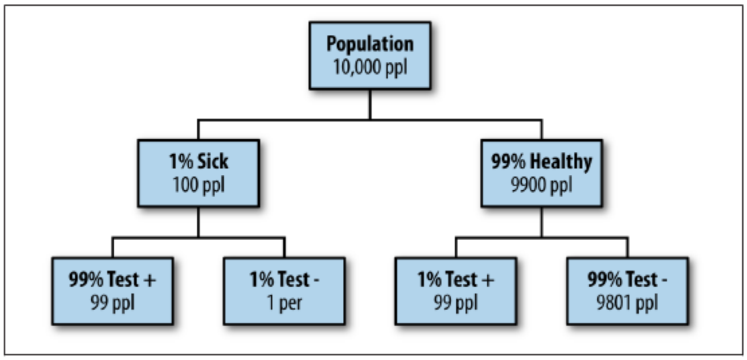
# Bayes Theorem

- Assume we are concentraing on a rare disease, where **1%** of the population is infected.
- We have a medical test to diagnose the disease. The test's performance is known as:
  - 99% of sick patients test positive
  - 99% of healthy patients test negative
- Given that a patient tests positive, what is the probability that the patient is actually sick?

Note: Please pay attention to the given information (i.e. *prior*) and desired information.

Question: What would be your estimate?

# Bayes Theorem




17

# Bayes Theorem with Total Probability

If $E_1$, $E_2$, … $E_k$ are $k$ mutually exclusive and exhaustive events and $B$ is any event,

$$P(E_1 \mid B) = \frac{P(B \mid E_1)P(E_1)}{P(B \mid E_1)P(E_1) + P(B \mid E_2)P(E_2) + \ldots + P(B \mid E_k)P(E_k)}$$

where $P(B) > 0$

Note : Numerator expression is always one of the terms in the sum of the denominator.

18

# Example: Bayesian Network

It is noted by Company X, a PC manufacturer, that when a customer calls the call center for filing a complaint about her PC, the reason is categorized into three: Hardware $P(H)$ = 0.1, software $P(S)$ = 0.6, and other $P(O)$ = 0.3.

Also, it is given that $P(F \mid H)$ = 0.9, $P(F \mid S)$ = 0.2, and $P(F \mid O)$ = 0.5.

If a failure occurs, determine if it's most likely due to hardware, software, or other.

Question-1: Why could this be important for operational purposes?

Question-2: What do you think is the answer? $H$, $S$ or $O$?

$$P(F) = P(F \mid H)P(H) + P(F \mid S)P(S) + P(F \mid O)P(O)$$
$$= 0.9(0.1) + 0.2(0.6) + 0.5(0.3) = 0.36$$
$$P(H \mid F) = \frac{P(F \mid H)\cdot P(H)}{P(F)} = \frac{0.9 \cdot 0.1}{0.36} = 0.250$$
$$P(S \mid F) = \frac{P(F \mid S)\cdot P(S)}{P(F)} = \frac{0.2 \cdot 0.6}{0.36} = 0.333$$
$$P(O \mid F) = \frac{P(F \mid O)\cdot P(O)}{P(F)} = \frac{0.5 \cdot 0.3}{0.36} = 0.417$$

Note that the conditionals given failure add to 1. Because $P(O \mid F)$ is largest, the most likely cause of the problem is in the *other* category.



# Classification: Definition

- Given a collection of records (*training set* )
  - Each record contains a set of *attributes*, one of the attributes is the *class (i.e. the "output")*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

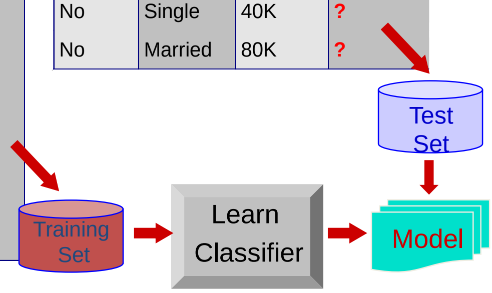
# Classification: Application

- Customer Attrition/Churn:
  - Goal: To predict whether a customer is likely to be lost to a competitor.
  - Approach:
    - Use detailed record of transactions with each of the past and present customers, to find attributes.
      - How often the customer uses mobile phone, does customer pay on-time, financial status, marital status, etc.
    - Label the customers as churner or non-churner.
    - Find a model for churn.

# Classification Example

| *ID* | Late Payment | Marital Status | Income | Defaulted |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | **No** |
| 2 | No | Married | 100K | **No** |
| 3 | No | Single | 70K | **No** |
| 4 | Yes | Married | 120K | **No** |
| 5 | No | Divorced | 95K | **Yes** |
| 6 | No | Married | 60K | **No** |
| 7 | Yes | Divorced | 220K | **No** |
| 8 | No | Single | 85K | **Yes** |
| 9 | No | Married | 75K | **No** |
| 10 | No | Single | 90K | **Yes** |

| Late Payme nt | Marital Status | Income | Defaulted |
|---|---|---|---|
| No | Single | 75K | ? |
| Yes | Married | 50K | ? |
| No | Married | 150K | ? |
| Yes | Divorced | 90K | ? |
| No | Single | 40K | ? |
| No | Married | 80K | ? |



# The weather problem

| Status | Levels |
|---|---|
| Outlook | {sunny, overcast, rainy} |
| Temperature | {hot, mild, cool} |
| Humidity | {high, normal} |
| Wind | {true, false} |

How many possible combinations?

A representative set of combinations is listed

| Outlook | Temp | Humidity | Windy | Play |
|---|---|---|---|---|
| Sunny | Hot | High | F | N |
| Sunny | Hot | High | T | N |
| Overcast | Hot | High | F | Y |
| Rainy | Mild | High | F | Y |
| Rainy | Cool | Normal | F | Y |
| Rainy | Cool | Normal | T | N |
| Overcast | Cool | Normal | T | Y |
| Sunny | Mild | High | F | N |
| Sunny | Cool | Normal | F | Y |
| Rainy | Mild | Normal | F | Y |
| Sunny | Mild | Normal | T | Y |
| Overcast | Mild | High | T | Y |
| Overcast | Hot | Normal | F | Y |
| Rainy | Mild | High | T | N |



# The weather problem

| Outlook | Temp | Humidity | Windy | Play |
|---|---|---|---|---|
| Sunny | Hot | High | F | N |
| Sunny | Hot | High | T | N |
| Overcast | Hot | High | F | Y |
| Rainy | Mild | High | F | Y |
| Rainy | Cool | Normal | F | Y |
| Rainy | Cool | Normal | T | N |
| Overcast | Cool | Normal | T | Y |
| Sunny | Mild | High | F | N |
| Sunny | Cool | Normal | F | Y |
| Rainy | Mild | Normal | F | Y |
| Sunny | Mild | Normal | T | Y |
| Overcast | Mild | High | T | Y |
| Overcast | Hot | Normal | F | Y |
| Rainy | Mild | High | T | N |

A set of rules learned from this information - not necessarily a very good one - might look as follows:

```
If outlook = sunny and humidity = high then play = no
If outlook = rainy and windy = true     then play = no
If outlook = overcast                   then play = yes
If humidity = normal                    then play = yes
If none of the above                    then play = yes
```

# The weather problem

| Outlook | Temp | Humidity | Windy | Play |
|---|---|---|---|---|
| Sunny | Hot | High | F | N |
| Sunny | Hot | High | T | N |
| Overcast | Hot | High | F | Y |
| Rainy | Mild | High | F | Y |
| Rainy | Cool | Normal | F | Y |
| Rainy | Cool | Normal | T | N |
| Overcast | Cool | Normal | T | Y |
| Sunny | Mild | High | F | N |
| Sunny | Cool | Normal | F | Y |
| Rainy | Mild | Normal | F | Y |
| Sunny | Mild | Normal | T | Y |
| Overcast | Mild | High | T | Y |
| Overcast | Hot | Normal | F | Y |
| Rainy | Mild | High | T | N |

- These rules are meant to be interpreted in order: the first one, then if it doesn't apply the second, and so on.
- A set of rules that are intended to be interpreted in sequence is called a *decision list.*
- Interpreted as a decision list, the rules correctly classify all of the examples in the table, whereas taken individually, out of context, some of the rules are incorrect. For example, the rule

  *if humidity = normal then play = yes*

  gets one of the examples wrong.

```
If outlook = sunny and humidity = high then play = no
If outlook = rainy and windy = true    then play = no
If outlook = overcast                  then play = yes
If humidity = normal                   then play = yes
If none of the above                   then play = yes
```

25

# A note: Classification vs. Association Rules

- The rules we have seen so far are *classification rules:* they predict the classification of the example in terms of whether to play or not.
- It is equally possible to disregard the classification and just look for any rules that strongly associate different attribute values. These are called *association rules.*
- Many association rules can be derived from the weather data.

Some association rules are as follows:

```
If temperature = cool                     then humidity = normal
If humidity = normal and windy = false    then play = yes
If outlook = sunny and play = no          then humidity = high
If windy = false and play = no            then outlook = sunny
                                               and humidity = high
```

26

## Simple probabilistic modeling for Classification

- Two assumptions: Attributes are
  - *equally important*
  - *statistically independent* (given the class value)
    - This means knowing the value of one attribute tells us nothing about the value of another takes on (if the class is known)
- Independence assumption is almost never correct!
- But … this scheme often works surprisingly well in practice
- The scheme is easy to implement in a program and very fast
- It is known as *Naïve Bayes*
- This method goes by the name of *Naïve Bayes*, because it's based on Bayes' rule and "*naïvely*" assumes independence

## Can combine probabilities using Bayes's rule

- Probability of an event *H* given observed evidence *E:*

$$P(H \mid E) = P(E \mid H)P(H) / P(E)$$

- *A priori* probability of *H* : $P(H)$
  - Probability of event *before* evidence is seen
- *A posteriori* probability of *H* : $P(H \mid E)$
  - Probability of event *after* evidence is seen
- Classification learning: what is the probability of the class given an instance?
  - Evidence *E* = instance's non-class attribute values
  - Event *H* = class value of instance
- Naïve assumption: evidence splits into parts (i.e., attributes) that are conditionally *independent*
- This means, given *n* attributes, we can write Bayes' rule using a product of per-attribute probabilities:

$$P(H \mid E) = P(E_1 \mid H)P(E_3 \mid H)...P(E_n \mid H)P(H) / P(E)$$

28

## Weather data example

| Outlook | Temp. | Humidity | Windy | Play |
|---|---|---|---|---|
| Sunny | Cool | High | True | ? |

$$P(yes \mid E) = P(Outlook = Sunny \mid yes)$$
$$P(Temperature = Cool \mid yes)$$
$$P(Humidity = High \mid yes)$$
$$P(Windy = True \mid yes)$$
$$P(yes) / P(E)$$

$$P(no \mid E) = P(Outlook = Sunny \mid no)$$
$$P(Temperature = Cool \mid no)$$
$$P(Humidity = High \mid no)$$
$$P(Windy = True \mid no)$$
$$P(no) / P(E)$$

29

## Weather data: Counts & Probabilities

| Outlook | | | Temperature | | | Humidity | | | Windy | | | Play | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Yes* | *No* | | *Yes* | *No* | | *Yes* | *No* | | *Yes* | *No* | *Yes* | *No* |
| Sunny | 2 | 3 | Hot | 2 | 2 | High | 3 | 4 | False | 6 | 2 | 9 | 5 |
| Overcast | 4 | 0 | Mild | 4 | 2 | Normal | 6 | 1 | True | 3 | 3 | | |
| Rainy | 3 | 2 | Cool | 3 | 1 | | | | | | | | |
| Sunny | 2/9 | 3/5 | Hot | 2/9 | 2/5 | High | 3/9 | 4/5 | False | 6/9 | 2/5 | 9/14 | 5/14 |
| Overcast | 4/9 | 0/5 | Mild | 4/9 | 2/5 | Normal | 6/9 | 1/5 | True | 3/9 | 3/5 | | |
| Rainy | 3/9 | 2/5 | Cool | 3/9 | 1/5 | | | | | | | | |

| Outlook | Temp | Humidity | Windy | Play |
|---|---|---|---|---|
| Sunny | Hot | High | False | No |
| Sunny | Hot | High | True | No |
| Overcast | Hot | High | False | Yes |
| Rainy | Mild | High | False | Yes |
| Rainy | Cool | Normal | False | Yes |
| Rainy | Cool | Normal | True | No |
| Overcast | Cool | Normal | True | Yes |
| Sunny | Mild | High | False | No |
| Sunny | Cool | Normal | False | Yes |
| Rainy | Mild | Normal | False | Yes |
| Sunny | Mild | Normal | True | Yes |
| Overcast | Mild | High | True | Yes |
| Overcast | Hot | Normal | False | Yes |
| Rainy | Mild | High | True | No |

## Weather data: Counts & Probabilities

| Outlook | | | Temperature | | | Humidity | | | Windy | | | Play | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Yes* | *No* | | *Yes* | *No* | | *Yes* | *No* | | *Yes* | *No* | *Yes* | *No* |
| Sunny | 2 | 3 | Hot | 2 | 2 | High | 3 | 4 | False | 6 | 2 | 9 | 5 |
| Overcast | 4 | 0 | Mild | 4 | 2 | Normal | 6 | 1 | True | 3 | 3 | | |
| Rainy | 3 | 2 | Cool | 3 | 1 | | | | | | | | |
| Sunny | 2/9 | 3/5 | Hot | 2/9 | 2/5 | High | 3/9 | 4/5 | False | 6/9 | 2/5 | 9/ 14 | 5/ 14 |
| Overcast | 4/9 | 0/5 | Mild | 4/9 | 2/5 | Normal | 6/9 | 1/5 | True | 3/9 | 3/5 | | |
| Rainy | 3/9 | 2/5 | Cool | 3/9 | 1/5 | | | | | | | | |

- A new day:

| Outlook | Temp. | Humidity | Windy | Play |
|---|---|---|---|---|
| Sunny | Cool | High | True | ? |

Likelihood of the two classes

For "yes" = 2/9 × 3/9 × 3/9 × 3/9 × 9/14 = 0.0053

For "no" = 3/5 × 1/5 × 4/5 × 3/5 × 5/14 = 0.0206

Conversion into a probability by normalization:

P("yes") = 0.0053 / (0.0053 + 0.0206) = 0.205

P("no") = 0.0206 / (0.0053 + 0.0206) = 0.795

31

## The "zero-frequency problem"

- What if an attribute value does not occur with every class value?
  (e.g., "Humidity = high" for class "yes")
  - Probability will be zero: $P(Humidity = High \mid yes) = 0$
  - *A posteriori* probability will also be zero: $P(yes \mid E) = 0$
    (Regardless of how likely the other values are!)
- Remedy?

- Remedy: add 1 to the count for every attribute value-class combination (*Laplace estimator)*
- Result: probabilities will never be zero
- Additional advantage: stabilizes probability estimates computed from small samples of data

32

## Modified probability estimates

- In some cases adding a constant different from 1 might be more appropriate
- Example: attribute *outlook* for class *yes*

$$\frac{2+\mu/3}{9+\mu} \qquad \frac{4+\mu/3}{9+\mu} \qquad \frac{3+\mu/3}{9+\mu}$$

***Sunny*** ***Overcast*** ***Rainy***

- Weights don't need to be equal (but they must sum to 1)

$$\frac{2+\mu p_1}{9+\mu} \qquad \frac{4+\mu p_2}{9+\mu} \qquad \frac{3+\mu p_3}{9+\mu}$$



## "Missing values" problem

- Training: instance is not included in frequency count for attribute value-class combination
- Classification: attribute will be omitted from calculation
- Example:

| Outlook | Temp. | Humidity | Windy | Play |
|---|---|---|---|---|
| ? | Cool | High | True | ? |

Likelihood of "yes" = 3/9 × 3/9 × 3/9 × 9/14 = 0.0238

Likelihood of "no" = 1/5 × 4/5 × 3/5 × 5/14 = 0.0343

P("yes") = 0.0238 / (0.0238 + 0.0343) = 41%

P("no") = 0.0343 / (0.0238 + 0.0343) = 59%

# Multinomial naïve Bayes I

- Version of naïve Bayes used for document classification using *bag of words* model
- $n_1, n_2, \ldots, n_k$: number of times word *i* occurs in the document
- $P_1, P_2, \ldots, P_k$: probability of obtaining word *i* when sampling from documents in class *H*
- Probability of observing a particular document *E* given probabilities class *H* (based on *multinomial distribution*):

$$P(E|H) = N! \times \prod_{i=1}^{k} \frac{P_i^{n_i}}{n_i!}$$

- Note that this expression ignores the probability of generating a document of the right length
  - This probability is assumed to be constant for all classes

35

# Multinomial naïve Bayes II

- Suppose dictionary has two words, *yellow* and *blue*
- Suppose P(*yellow* | *H*) = 75% and P(*blue* | *H*) = 25%
- Suppose *E* is the document "*blue yellow blue"*
- Probability of observing document:

$$P(\{blueyellowblue\} \mid H) = 3! \times \frac{0.75^1}{1!} \times \frac{0.25^2}{2!} = \frac{27}{64}$$

Suppose there is another class *H*' that has P(*yellow* | *H'*) = 10% and P(*blue*| *H'*) = 90%:

$$P(\{blueyellowblue\} \mid H) = 3! \times \frac{0.1^1}{1!} \times \frac{0.9^2}{2!} = \frac{243}{1000}$$

- Need to take prior probability of class into account to make the final classification using Bayes' rule
- Factorials do not actually need to be computed: they drop out
- Underflows can be prevented by using logarithms

36

# Naïve Bayes: discussion

- Naïve Bayes works surprisingly well even if independence assumption is clearly violated
- Why? Because classification does not require accurate probability estimates *as long as maximum probability is assigned to the correct class*
- However: adding too many redundant attributes will cause problems (e.g., identical attributes)
- Note also: many numeric attributes are not normally distributed (*kernel density estimators* can be used instead)