Inference using the joint distribution

<table>
<thead>
<tr>
<th></th>
<th>ache</th>
<th>¬ache</th>
</tr>
</thead>
<tbody>
<tr>
<td>cavity</td>
<td>0.4</td>
<td>0.1</td>
</tr>
<tr>
<td>¬cavity</td>
<td>0.1</td>
<td>0.4</td>
</tr>
</tbody>
</table>

\[ P(\text{cavity}) = P(\text{cavity, ache}) + P(\text{cavity, ¬ache}) \]

Notation

Let \( Y, Z \) denote sets of random variables

\[ Y = \{ Y_1, Y_2, Y_3 \}; \quad P(Y) = P(Y_1, Y_2, Y_3) \]

\[ Z = \{ Z_1, Z_2 \}; \quad P(Z) = P(Z_1, Z_2) \]

\[ P(Y \cup Z) = P(Y_1, Y_2, Y_3, Z_1, Z_2) = P(Y, Z) \]

\[ P(Y \mid Z) = \frac{P(Y \cup Z)}{P(Z)} \]

Note the overloading of \( P \) and an unfortunate consequence of the set notation.
Marginalization

Let \( Y, Z \) denote sets of random variables.
\[ P(Y) = \sum_{x} P(Y, x) \] where the summation is over all assignment of values to random variables in \( Z \).

Similarly, \( P(Y) = \sum_{x} P(Y \mid x) P(x) \).

Example: \( Y = \{y_1, y_2, y_3\} \), \( Z = \{z_1, z_2\} \)
Suppose all random variables are binary.
The joint distribution over the variables in \( Z \cup Y \) has \( 2^5 \) entries.
Marginalization over \( Y \) results in a joint distribution over the variables in \( Z \) yielding a table of \( 2^3 \) entries.

Independence and Conditional Independence

Let \((E, P)\) be a probability space. Let \( A_1, A_2 \subseteq E \). We say that the events \( A_1 \) and \( A_2 \) are independent if
\[ P(A_1 \land A_2) = P(A_1)P(A_2) \]

\[ \iff \begin{cases} P(A_1) \neq 0, P(A_2) \neq 0, & P(A_1 \mid A_2) = P(A_1) \text{ and } P(A_2 \mid A_1) = P(A_2) \iff A_1 \text{ and } A_2 \text{ are independent} \end{cases} \]

\[ \text{If } P(A_1) = 0 \text{ or } P(A_2) = 0 \text{ (or both), } P(A_1 \land A_2) = 0 \]

Independence and Conditional Independence

If for every subset \( B = \{B_1, \ldots, B_n\} \) obtained by selecting \( k \) elements of \( A = \{A_1, \ldots, A_n\} \) \((1 \leq k \leq n)\) if we have
\[ P(B_1, \ldots, B_n \mid C) = \prod_{j=1}^{n} P(B_j \mid C) \] we say that \( A_1, \ldots, A_n \) are mutually independent given \( C \).
Conditional Independence

*X* is conditionally independent of *Y* given *Z* if the probability distribution governing *X* is independent of the value of *Y* given the value of *Z*:

\[ P(X|Y, Z) = P(X|Z) \]

That is, if

\[(Y, x, y, z) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k) \]

Conditional Independence

Thunder is independent of Rain given Lightning

\[
P(\text{Thunder} = 1|\text{Rain} = 1, \text{Lightning} = 1) = P(\text{Thunder} = 1|\text{Lightning} = 1) = P(\text{Thunder} = 1|\text{Rain} = 0, \text{Lightning} = 1) = P(\text{Thunder} = 1|\text{Rain} = 0, \text{Lightning} = 0)
\]

\[
P(\text{Thunder} = 1|\text{Rain} = 0, \text{Lightning} = 0) = P(\text{Thunder} = 0|\text{Lightning} = 1) = P(\text{Thunder} = 0|\text{Rain} = 0, \text{Lightning} = 1)
\]

\[
P(\text{Thuder} = 0|\text{Rain} = 1, \text{Lightning} = 0) = P(\text{Thunder} = 0|\text{Lightning} = 0)
\]

\[
P(\text{Thunder} = 0|\text{Rain} = 0, \text{Lightning} = 0) = P(\text{Thunder} = 0|\text{Rain} = 0, \text{Lightning} = 0)
\]

Independence and Conditional Independence

Let \( Z_1, ..., Z_k \) and \( W \) be pairwise disjoint sets of random variables on a given event space. \( Z_1, ..., Z_k \) are mutually independent given \( W \) if

\[
P(\bigcup_{i=1}^{k} Z_i | W) = \prod_{i=1}^{k} P(Z_i | W)
\]

\[
P(Z_1 | Z_2 \cup W) = P(Z_1 | W) \text{ if } Z_1 \text{ and } Z_2 \text{ are independent.}
\]

Note that these represent sets of equations, for all possible value assignments to random variables.
Independence Properties of Random Variables

Let $W, X, Y, Z$ be pairwise disjoint sets of random variables on a given event space.
Let $I(X, Y, Z)$ denote that $X$ and $Z$ are independent given $Y$.
That is, $P(X \cup Y | Z) \neq P(X | Z) P(Y | Z)$ or $P(X \cup Y | Z) \neq P(X | Z) P(Y | Z)$
Then:

a. $I(X, Z, Y) \Rightarrow I(Y, Z, X)$
b. $I(X, Z, Y \cup W) \Rightarrow I(X, Z, Y)$
c. $I(X, Z, Y \cup W) \Rightarrow I(X, Z \cup W, Y)$
d. $I(X, Z, Y) \wedge I(X, Z \cup Y, W) \Rightarrow I(X, Z, Y \cup W)$
Proof: Follows from definition of independence.

Expectation and Variance

Let $X : \mathbb{E} \rightarrow \mathbb{R}$ be a random variable on a finite probability space $(\mathbb{E}, P)$ and $B \subseteq \mathbb{E}$.
The conditional expectation or expected value of $X$ given $B$ is
$$E(X | B) = \sum_{e \in B} P(e | B) X(e) = \frac{1}{P(B)} \sum_{e \in \mathbb{E}} P(e | B) X(e)$$
The variance of $X$ given $B$ is given by
$$\text{Var}(X | B) = E((X - E(X | B))^2 | B)$$
The unconditional expectation and variance correspond to the case $B = \mathbb{E}$ in which case we simply drop "$| B$".

Conditional expectation of random variables

Expectation of a random variable $X$ conditioned on a random variable $Y$ is $E(X | Y) = \sum_{e \in \mathbb{E}} P(e | Y) X(e)$
Note that this denotes a set of equations for possible values of $Y$.
The definitions can be extended to the case where $X$ and $Y$ are replaced by sets of random variables.

Example

| $P(X=0 | Y=0)$ | $P(X=1 | Y=0)$ | $P(X=0 | Y=1)$ | $P(X=1 | Y=1)$ |
|----------------|----------------|----------------|----------------|
| 0.6            | 0.4            | 0.7            | 0.3            |

$$E(X | Y = 0) = P(X = 0 | Y = 0) \times 0 + P(X = 1 | Y = 0) \times 1 = 0.4$$
$$E(X | Y = 1) = 0.7$$
Properties of Expectation and Variance

Let \( X, X_1, \ldots, X_n \) be random variables and \( a, b, c, \ldots \) be real numbers.

\[ \text{Var}(X | \theta) = E(X^2 | \theta) - E(X | \theta)^2 \]

If \( X \) has mean \( \mu \) and variance \( \sigma^2 \), then \( aX + b \) has mean \((a\mu + b)\) and variance \(a^2\sigma^2\).

For any \( c, \theta \) and \( X, X_1, \ldots, X_n \),

\[ E \left( \sum c_i X_i | \theta \right) = \sum c_i E(X_i | \theta) \]

If \( \forall i \neq j \) \( X_i \), and \( X_j \), are independent given \( \theta \), then

\[ \text{Var} \left( \sum c_i X_i | \theta \right) = \sum c_i^2 \text{Var}(X_i | \theta) \]

Proof of these results is left as an exercise.

Learning as Bayesian Inference

Probability is the logic of Science (Jaynes)

Bayesian (subjective) probability provides a basis for updating beliefs based on evidence.

By updating beliefs about hypotheses based on data, we can learn about the world.

Bayesian framework provides a sound probabilistic basis for understanding many learning algorithms and designing new algorithms.

Bayesian framework provides several practical reasoning and learning algorithms.

Bayesian Classification

Consider the problem of classifying an instance \( X \) into one of two mutually exclusive classes \( \omega_1 \) or \( \omega_2 \):

\[ P(\omega_1 | X) = \text{probability of class } \omega_1 \text{ given the evidence } X \]

\[ P(\omega_2 | X) = \text{probability of class } \omega_2 \text{ given the evidence } X \]

What is the probability of error?

\[ P(\text{error} | X) = P(\omega_1 | X) \text{ if we choose } \omega_2, \]

\[ = P(\omega_2 | X) \text{ if we choose } \omega_1, \]
Bayesian Optimal Classification

To minimize classification error
Choose \( \omega_1 \) if \( P(\omega_1 | X) > P(\omega_2 | X) \)
Choose \( \omega_2 \) if \( P(\omega_2 | X) > P(\omega_1 | X) \)
which yields
\[
P(error | X) = \min[P(\omega_1 | X), P(\omega_2 | X)]
\]
We have:
\[
P(\omega_1 | X) = P(X | \omega_1) P(\omega_1);
P(\omega_2 | X) = P(X | \omega_2) P(\omega_2)
\]

Bayes Optimal Classification

Classification rule that guarantees minimum error:
Choose \( \omega_1 \) if \( P(X | \omega_1) P(\omega_1) > P(X | \omega_2) P(\omega_2) \)
Choose \( \omega_2 \) if \( P(X | \omega_2) P(\omega_2) > P(X | \omega_1) P(\omega_1) \)

If \( P(X | \omega_1) = P(X | \omega_2) \), classification depends entirely on \( P(\omega_1) \) and \( P(\omega_2) \)

If \( P(\omega_1) = P(\omega_2) \), classification depends entirely on \( P(X | \omega_1) \) and \( P(X | \omega_2) \)

Bayes classification rule combines the effect of the two terms optimally - so as to yield minimum error classification.
Generalization to multiple classes: \( c(X) = \arg \max_{\omega_i} P(\omega_i | X) \)

Optimality of Bayesian classifier

We can show that the Bayesian classifier is optimal in that it is guaranteed to minimize the probability of misclassification

(Proof given in class)
Minimum Risk Classification

Let $\lambda_i$ be the cost or risk associated with assigning an instance to class $o_i$ when the correct classification is $o_j$.

- $R(o_i | X) = \text{expected loss incurred in assigning } X \text{ to class } o_i$
- $R(o_i | X) = \lambda_{ij} \cdot P(o_i | X) + \lambda_{ji} \cdot P(o_j | X)$

Classification rule that guarantees minimum risk:
- Choose $o_1$ if $R(o_1 | X) < R(o_2 | X)$
- Choose $o_2$ if $R(o_1 | X) < R(o_2 | X)$
- Flip a coin otherwise

This classification rule can be shown to be optimal in that it is guaranteed to minimize the risk of misclassification.

Summary of Bayesian recipe for classification

- $\lambda_i$ be the cost or risk associated with assigning an instance to class $o_i$ when the correct classification is $o_j$.

Choose $o_1$ if $P(X|o_1) / P(X|o_2) > (\lambda_{ij} - \lambda_{ji}) / \lambda_{ji}$

Choose $o_2$ if $P(X|o_2) / P(X|o_1) > (\lambda_{ij} - \lambda_{ji}) / \lambda_{ij}$

Minimum error classification rule is a special case:

$\lambda_i = 0$ if $i = j$ and $\lambda_i = 1$ if $i \neq j$
Summary of Bayesian recipe for classification

The Bayesian recipe is simple, optimal, and in principle, straightforward to apply.
To use this recipe in practice, we need to know $P(X|\omega_i)$ and $P(\omega_i)$.
Because these probabilities are unknown, we need to estimate them from data – or learn them!
- $X$ is typically a high-dimensional vector
- Need to estimate $P(X|\omega_i)$ from limited data

Naïve Bayes Classifier

We can classify $X$ if we know $P(X|\omega_i)$.
How to learn $P(X|\omega_i)$?
One solution: Assume that the random variables in $X$ are conditionally independent given the class.
Result: Naïve Bayes classifier which performs optimally under certain assumptions.
A simple, practical learning algorithm grounded in Probability Theory.
When to use
- Attributes that describe instances are likely to be conditionally independent given classification.
- The data is insufficient to estimate all the probabilities reliably if we do not assume independence – which is often the case.

Successful applications
- Diagnosis
- Document Classification
- Protein Function Classification
- Prediction of protein-protein interfaces
- and many others…….
Conditional Independence

Let $Z_1, \ldots, Z_n$ and $W$ be random variables on a given event space. $Z_1, \ldots, Z_n$ are mutually independent given $W$ if

$$P(Z_1, Z_2, \ldots, Z_n | W) = \prod_{i=1}^n P(Z_i | W)$$

Note that these represent sets of equations, for all possible value assignments to random variables.

Implications of Independence

Suppose we have 5 Binary attributes and a binary class label $y$.

Without independence, in order to specify the joint distribution, we need to specify a probability for each possible assignment of values to each variable resulting in a table of size $2^n = 64$.

Suppose the features are independent given the class label $y$. We only need $5(2^2) = 20$ entries.

The reduction in the number of probabilities to be estimated is even more striking when $n$, the number of attributes is large – from $O(2^n)$ to $O(n)$.

Naive Bayes Classifier

Consider a discrete valued target function $f : \chi \rightarrow \Omega$ where an instance $X = (X_1, X_2, \ldots, X_n) \in \chi$ is described in terms of attribute values $X_1 = x_1$, $X_2 = x_2$, ..., $X_n = x_n$.

where $x_i \in \text{Domain}(X_i)$

$\omega_{\text{MAP}} = \arg\max_{\omega \in \Omega} P(\omega) \cdot P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | \omega)$

$= \arg\max_{\omega \in \Omega} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | \omega) P(\omega)$

$\omega_{\text{MAP}}$ is called the maximum a posteriori classification.
Naive Bayes Classifier

\[
\begin{align*}
\omega_{MAP} &= \arg\max_{\omega_i \in \Omega} P(\omega_i | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\
&= \arg\max_{\omega_i \in \Omega} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | \omega_i) P(\omega_i)
\end{align*}
\]

If the attributes are independent given the class, we have

\[
\begin{align*}
\omega_{MAP} &= \arg\max_{\omega_i \in \Omega} \prod_{j=1}^n P(X_j = x_j | \omega_i) P(\omega_i) \\
&= \omega_i_{MAP}
\end{align*}
\]

\[
\begin{align*}
\omega_{MAP} &= \arg\max_{\omega_i \in \Omega} \prod_{j=1}^n P(X_j = x_j | \omega_i)
\end{align*}
\]

Naive Bayes Learner

For each possible value \(\omega_i\) of \(\Omega\),

\[
\hat{P}(\Omega = \omega_i) \leftarrow \text{Estimate}(P(\Omega = \omega_i), D)
\]

For each possible value \(a_j\) of \(X_j\),

\[
\hat{P}(X_j = a_j | \omega_i) \leftarrow \text{Estimate}(P(X_j = a_j | \Omega = \omega_i), D)
\]

Classify a new instance \(X = (x_1, x_2, \ldots, x_n)\)

\[
c(X) = \arg\max_{\omega_i \in \Omega} \prod_{j=1}^n P(X_j = x_j | \omega_i)
\]

\text{Estimate} is a procedure for estimating the relevant probabilities from set of training examples

Estimating probabilities from data (discrete case)

Maximum likelihood estimation
Bayesian estimation
Maximum a posteriori estimation
Example: Binomial Experiment

When tossed, the thumbtack can land in one of two positions: Head or Tail.

We denote by \( \theta \) the (unknown) probability \( P(H) \).

Estimation task—

Given a sequence of toss samples \( x[1], x[2], \ldots, x[M] \) we want to estimate the probabilities \( P(H) = \theta \) and \( P(T) = 1 - \theta \).

Statistical parameter fitting

Consider samples \( x[1], x[2], \ldots, x[M] \) such that:

- The set of values that \( X \) can take is known.
- Each is sampled from the same distribution.
- Each is sampled independently of the rest.

The task is to find a parameter \( \Theta \) so that the data can be summarized by a probability \( P(x[j] | \Theta) \):

- The parameters depend on the given family of probability distributions: multinomial, Gaussian, Poisson, etc.
- We will focus first on binomial and then on multinomial distributions.
- The main ideas generalize to other distribution families.

The Likelihood Function

How good is a particular \( \theta \)?

It depends on how likely it is to generate the observed data.

The likelihood for the sequence \( H, T, T, H, H \) is:

\[ L(\theta : D) = P(D | \theta) = \prod_{i=1}^{5} P(x[i] | \theta) \]

\[ L(\theta : D) = \theta \cdot (1 - \theta) \cdot (1 - \theta) \cdot \theta \cdot 0 \]

\[ L(\theta : D) = \theta \cdot (1 - \theta) \cdot (1 - \theta) \cdot \theta \cdot 0 \]

\[ L(\theta : D) = \theta \cdot (1 - \theta) \cdot (1 - \theta) \cdot \theta \cdot 0 \]

\[ L(\theta : D) = \theta \cdot (1 - \theta) \cdot (1 - \theta) \cdot \theta \cdot 0 \]
**Likelihood function**

The likelihood function \( L(\theta : D) \) provides a measure of relative preferences for various values of the parameter \( \theta \) given a collection of observations \( D \) drawn from a distribution that is parameterized by fixed but unknown \( \theta \).

\[ L(\theta : D) \text{ is the probability of the observed data } D \text{ considered as a function of } \theta. \]

Suppose data \( D \) is 5 heads out of 8 tosses. What is the likelihood function assuming that the observations were generated by a binomial distribution with an unknown but fixed parameter \( \theta \)?

\[
\binom{8}{5} \theta^5 (1-\theta)^3
\]

**Sufficient Statistics**

To compute the likelihood in the thumbtack example we only require \( N_H \) and \( N_T \) (the number of heads and the number of tails).

\[
L(0 : D) = 0^{N_H} \cdot (1-0)^{N_T}
\]

\( N_H \) and \( N_T \) are **sufficient statistics** for the parameter \( \theta \) that specifies the binomial distribution.

A statistic is simply a function of the data.

A **sufficient statistic** \( s \) for a parameter \( \theta \) is a function that summarizes from the data \( D \), the relevant information \( s(D) \) needed to compute the likelihood \( L(\theta : D) \).

If \( s \) is a sufficient statistic for \( s(D) = s(D') \), then \( L(\theta : D) = L(\theta : D') \).

**Maximum Likelihood Estimation**

**Main Idea:** Learn parameters that maximize the likelihood function.

Maximum likelihood estimation is
- Intuitively appealing
- One of the most commonly used estimators in statistics
- Assumes that the parameter to be estimated is fixed, but unknown
Example: MLE for Binomial Data

Applying the MLE principle we get

\[ \hat{\theta} = \frac{N_H}{N_H + N_T} \]

(Why?)

Example: 
\[ (N_H, N_T) = (3, 2) \]

ML estimate is \( \frac{3}{5} = 0.6 \)

The likelihood is positive for all legitimate values of \( \theta \)

So maximizing the likelihood is equivalent to maximizing its logarithm i.e. log likelihood

\[ \frac{\partial}{\partial \theta} \log L(\theta; D) = \frac{N_H - N_T}{1 - \theta} = 0 \]

Note that the likelihood is indeed maximized at \( \theta = \theta_{ML} \) because in the neighborhood of \( \theta_{ML} \), the value of the likelihood is smaller than it is at \( \theta = \theta_{ML} \)

Maximum and curvature of likelihood around the maximum

At the maximum, the derivative of the log likelihood is zero
At the maximum, the second derivative is negative.

The curvature of the log likelihood is defined as

\[ I(\theta) = -\frac{\partial^2}{\partial \theta^2} \log L(\theta; D) \]

Large observed curvature \( I(\theta_{ML}) \) at \( \theta = \theta_{ML} \) is associated with a sharp peak, intuitively indicating less uncertainty about the maximum likelihood estimate \( \theta_{ML} \) is called the Fisher information
**Maximum Likelihood Estimate**

ML estimate can be shown to be:
- **Asymptotically unbiased**
  \[ \lim_{n \to \infty} E[ \theta_{ML} ] = \theta_{true} \]
- **Asymptotically consistent** - converges to the true value as the number of examples approaches infinity
  \[ \lim_{n \to \infty} P( | \theta_{ML} - \theta_{true} | \leq \varepsilon ) = 1 \]
  \[ \lim_{n \to \infty} E[ | \theta_{ML} - \theta_{true} | ] = 0 \]
- **Asymptotically efficient** - achieves the lowest variance that any estimate can achieve for a training set of a certain size (satisfies the Cramer-Rao bound)

**Maximum Likelihood Estimate**

ML estimate can be shown to be representationally invariant:
- If \( \theta_{ML} \) is an ML estimate of \( \theta \), and \( g(\theta) \) is a function of \( \theta \), then \( g(\theta_{ML}) \) is an ML estimate of \( g(\theta) \)

When the number of samples is large, the probability distribution of \( \theta_{ML} \) has Gaussian distribution with mean \( \theta_{true} \) (the actual value of the parameter) - a consequence of the central limit theorem that says that a random variable which is a sum of a large number of random variables has a Gaussian distribution and the fact that the ML estimate is related to the sum of random variables.

We can use the likelihood ratio to reject the null hypothesis corresponding to \( \theta = \theta_0 \) as unsupported by data if the ratio of the likelihoods evaluated at \( \theta_0 \) and at \( \theta_{ML} \) is small.
(The ratio can be calibrated when the likelihood function is approximately quadratic)

**Naïve Bayes Classifier**

We can define the likelihood for a Naïve Bayesian Classifier:

Let \( \Theta_j \) be the class conditional probabilities for class \( j \)
Let \( L_j \) be the corresponding likelihood

\( L_j(\Theta;D) = \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \prod \ projectile
Naive Bayes Classifier

Decomposition → Independent Estimation Problems

If the parameters for each family are decoupled via independence, then they can be estimated independently of each other.

From Binomial to Multinomial

Suppose a random variable \( X \) can take the values 1, 2, ..., \( K \). We want to learn the parameters \( \theta_1, \theta_2, ..., \theta_K \).

Sufficient statistics: \( N_1, N_2, ..., N_K \) - the number of times each outcome is observed.

Likelihood function

\[
L(\theta; D) = \prod_{k=1}^{K} \theta_k^{N_k}
\]

ML estimate

\[
\hat{\theta}_k = \frac{N_k}{\sum_i N_i}
\]

MLE estimates for Naive Bayes Classifiers

When we assume that \( P(X_i | C) \) is multinomial, we get the decomposition:

\[
L(\theta; D) = \prod_{i} P(x_i, [m_i]; [w_i] \cdot \Theta_i) = \prod_{i} \prod_{j} P(x_i | C_j; \Theta_j)^{N_i(C_j)} = \prod_{i} \prod_{j} \theta_j^{N_i(C_j)}
\]

For each class we get an independent multinomial estimation problem.

The MLE is

\[
\hat{\theta}_{i,j} = \frac{N(x_i, C_j)}{N(C_j)}
\]
Summary of Maximum Likelihood estimation

Define a *likelihood function* which is a measure of how likely it is that the observed data were generated from a probability distribution with a particular choice of parameters.

Select the parameters that maximize the likelihood.

In simple cases, ML estimate has a closed form solution.

In other cases, ML estimation may require numerical optimization.

**Problem with ML estimate** - assigns zero probability to unobserved values - can lead to difficulties when estimating from small samples.

**Question** - How would Naïve Bayes classifier behave if some of the class conditional probability estimates are zero?