Bayesian Learning
Multiclass Classification

M. R. Hasan
Readings

- Bishop: 2.2
- Murphy: 3.4
Learning Strategy

• We divide the learning techniques in two categories:
• Frequentist Learning
• Bayesian Learning
MLE’s Blind Spot

The black-swan paradox can be solved by resorting to **Bayesian learning.**
Bayesian Learning: Multinomial Variable

• So far we have discussed how to infer the probability that a coin comes up heads (binary variables)
• Now, we generalize these results to infer the probability that a dice with K sides comes up as face k.
• We model the outcome of a K-sided die toss by using multinomial variables.
• The methods we will study are widely used to analyze text data, biosequence data, etc.
Bayesian Learning: Multinomial Variable

• **Likelihood:**
• Suppose we observe N *dice rolls*, $D = \{x_1, ..., x_N\}$, where $x_i \in \{1, ..., K\}$.
• In each roll, $x_i$ will **show one of the K sides up**.
• So the variable is represented by a **K-dimensional vector** $x$ in which one of the elements $x_i$ equals to 1.
• All remaining elements are 0.
Bayesian Learning: Multinomial Variable

• For example, say we have a die with 6 faces.
• So, the variable can take $K = 6$ states.
• These states are represented by a *categorical variable* $x$.
• We represent $x$ by *one-hot encoding*.

<table>
<thead>
<tr>
<th>Bird</th>
<th>Robin</th>
<th>Meadowlark</th>
<th>Quail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robin</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Robin</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Meadowlark</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Quail</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Robin</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

One-hot encoding of the *categorical variable* “Bird” in 5 samples
Bayesian Learning: Multinomial Variable

- A particular observation of the variable happens to correspond to the state where \( x_3 = 1 \).
- Then, \( x \) will be represented using one-hot encoding:

\[
x = (0, 0, 1, 0, 0, 0, 0)^T
\]

The vector \( x \) satisfies:

\[
\sum_{i=k}^{K} x_k = 1
\]
Bayesian Learning: Multinomial Variable

• If we denote the probability \( x_i = 1 \) by the parameter \( \theta_i \), then the distribution of \( x \) is given by:

\[
p(\vec{x} | \vec{\theta}) = \prod_{k=1}^{K} \theta_k^{x_k}
\]

Here \( \theta = (\theta_1, \ldots, \theta_K)^T \) represents the probabilities and satisfy the following constraints:

\[
\theta_k \geq 1
\]

\[
\sum_k \theta_k = 1
\]

Example: \( K = 6 \) & \( x_3 = 1 \)

\[
x = (0, 0, 1, 0, 0, 0)^T
\]

Probability of getting the k’th side up \( (x_k = 1) \) is \( \theta_k \)
Bayesian Learning: Multinomial Variable

- The following distribution can be regarded as a generalization of the Bernoulli distribution to more than two outcomes.

\[ p(\tilde{x} \mid \tilde{\theta}) = \prod_{k=1}^{K} \theta_{k}^{x_{k}} \]

Probability of getting the k’th side up \((x_k = 1)\) is \(\theta_k\)

This distribution is normalized:

\[ \sum_{\tilde{x}} p(\tilde{x} \mid \tilde{\theta}) = \sum_{k=1}^{K} \theta_k = 1 \]

The expectation is:

\[ \mathbb{E}[\tilde{x} \mid \tilde{\theta}] = \sum_{\tilde{x}} p(\tilde{x} \mid \tilde{\theta}) \tilde{x} = (\theta_1, \ldots, \theta_N)^T = \tilde{\theta} \]

Example: \(K = 6\) & \(x_3 = 1\)

\[ x = (0, 0, 1, 0, 0, 0)^T \]
Multinomial Variable

Now we can write the expression for the likelihood function.

Suppose we observe N dice rolls, \( D = \{x_1, \ldots, x_N\} \), where \( x_i \in \{1, \ldots, K\} \).

If we assume the data is i.i.d., the likelihood function has the form:

\[
p(D | \tilde{\theta}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \theta_{nk}^{x_{nk}}
\]

\[
= (\theta_1^{x_1} \cdot \theta_2^{x_2} \cdots \theta_k^{x_k}) \cdot (\theta_1^{x_1} \cdot \theta_2^{x_2} \cdots \theta_k^{x_k}) \cdots (\theta_1^{x_1} \cdot \theta_2^{x_2} \cdots \theta_k^{x_k})
\]

\[
= (\theta_1^{x_1} \cdot \theta_1^{x_1} \cdots \theta_1^{x_1}) \cdot (\theta_2^{x_2} \cdot \theta_2^{x_2} \cdots \theta_2^{x_2}) \cdots (\theta_k^{x_k} \cdot \theta_k^{x_k} \cdots \theta_k^{x_k})
\]
Bayesian Learning: Multinomial Variable

\[ p(D | \vec{\theta}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \theta_{k}^{x_{nk}} \]

\[ = \left( \theta_{1}^{x_{1}} \cdot \theta_{2}^{x_{2}} \cdots \theta_{k}^{x_{k}} \right) \cdot \left( \theta_{1}^{x_{1}} \cdot \theta_{2}^{x_{2}} \cdots \theta_{k}^{x_{k}} \right) \cdots \left( \theta_{1}^{x_{1}} \cdot \theta_{2}^{x_{2}} \cdots \theta_{k}^{x_{k}} \right) \]

\[ = \left( \theta_{1}^{x_{1}} \cdot \theta_{1}^{x_{1}} \cdots \theta_{1}^{x_{1}} \right) \cdot \left( \theta_{2}^{x_{2}} \cdot \theta_{2}^{x_{2}} \cdots \theta_{2}^{x_{2}} \right) \cdots \left( \theta_{k}^{x_{k}} \cdot \theta_{k}^{x_{k}} \cdots \theta_{k}^{x_{k}} \right) \]

\[ = \left( \theta_{1}^{\sum_{n} x_{n1}} \right) \cdot \left( \theta_{2}^{\sum_{n} x_{n2}} \right) \cdots \left( \theta_{k}^{\sum_{n} x_{nk}} \right) \]

\[ = \prod_{k=1}^{K} \theta_{k}^{\sum_{n} x_{nk}} \]

Ex: No. of times \( x_{1} \) occurred in \( N \) trials

\( \sum_{n} x_{nk} \): the number of times event \( k \) occurred in \( N \) trials

Example: \( K = 6 \) & \( x_{3} = 1 \)

\[ x = (0, 0, 1, 0, 0, 0)^{T} \]
Bayesian Learning: Multinomial Variable

- Notice that the likelihood function depends on the N data points only through the K quantities:

\[
p(D \mid \hat{\theta}) = \prod_{k=1}^{K} \theta_k^{\sum_{n} x_{nk}}
\]

Denote \( N_k \) to represent the number of times event \( k \) occurred.

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

\( N_k \) is called sufficient statistic for this distribution.

Example: \( K = 6 \) & \( x_3 = 1 \)

\[
x = (0, 0, 1, 0, 0, 0)^T
\]
Bayesian Learning: Multinomial Variable

• We can consider the joint distribution of the quantities $N_1$, \ldots, $N_K$, conditioned on the parameters $\theta$ and on the total number $N$ of observations.

• Hence, from $p(D \mid \theta)$ we can write the expression for the multinomial distribution:

$$p(D \mid \hat{\theta}) = \prod_{k=1}^{K} \theta_k^{N_k}$$

\[ N_k = \sum_n x_{nk}: \text{the number of times event } k \text{ occurred in } N \text{ trials} \]

$$\text{Mult}(N_1, N_2, \ldots, N_K \mid \hat{\theta}, N) = \binom{N}{N_1 N_2 \ldots N_K} \prod_{k=1}^{K} \theta_k^{N_k}$$
Bayesian Learning: Multinomial Variable

- The normalization coefficient is the **number of ways of partitioning N objects into K groups** of size $N_1, \ldots, N_K$:

\[
\binom{N}{N_1 N_2, \ldots, N_K} = \frac{N!}{N_1! N_2! \ldots N_K!}
\]

$N_k$ variables are subject to the constraint:

\[
\sum_{k=1}^{K} N_k = N
\]

\[
Mult(N_1, N_2, \ldots, N_k | \theta, N) = \binom{N}{N_1 N_2 \ldots N_k} \prod_{k=1}^{K} \theta_k^{N_k}
\]
Bayesian Learning: Multinomial Variable

- Now that we have an expression for the likelihood function, let’s determine the **prior distributions** for the parameters \( \{\theta_k\} \) of the multinomial distribution (at the bottom).
- **By inspection** of the form of the multinomial distribution, we see that the **conjugate prior**:

\[
p(\hat{\theta} | \tilde{\alpha}) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}
\]

Where

\[
\sum_{k} \theta_k = 1
\]

\[
0 \leq \theta_k \leq 1
\]

\[
Mult(N_1, N_2, \ldots, N_k | \hat{\theta}, N) = \binom{N}{N_1 N_2 \ldots N_k} \prod_{k=1}^{K} \theta_k^{N_k}
\]
Bayesian Learning: Multinomial Variable

• Here \( \alpha_1, \ldots, \alpha_K \) are the parameters of the distribution, and \( \vec{\alpha} \) denotes \((\alpha_1, \ldots, \alpha_K)^T\).

• The normalized form for this distribution is called the Dirichlet distribution.

\[
p(\vec{\theta} | \vec{\alpha}) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}
\]

\[
Dir(\vec{\theta} | \vec{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}
\]

Here \( \Gamma(x) \) is a gamma function

\[
\Gamma(x) \equiv \int_0^{\infty} u^{x-1} e^{-u} \, du.
\]

\[\alpha_0 = \sum_{k=1}^{K} \alpha_k\]
Bayesian Learning: Multinomial Variable

• Note that, Dirichlet distribution is a distribution of distributions $\theta_k$.

• Also because of the summation constraint (all probabilities sum to 1), the distribution over the space of the $\{\theta_k\}$ is confined to a simplex of dimensionality $K - 1$.

\[
\sum_{k} \theta_k = 1 \quad \quad 0 \leq \theta_k \leq 1
\]

\[
p(\tilde{\theta} | \tilde{\alpha}) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k-1}
\]
Bayesian Learning: Multinomial Variable

• **Simplex**:  
  • If we roll a K-sided dice, the probability of landing on each side is represented by a probability distribution (pmf) $\theta_k$ over a K-dimensional space.  
  • For example, for a fair die (6 sided: $k = 1:6$), $\theta_k = 1/6$  
  • The probabilities are summed to 1 (for all K-dimensions) and $0 \leq \theta_k \leq 1$

A simplex represents the plot of these probabilities.
Bayesian Learning: Multinomial Variable

• Simplex:
• For example, if $K = 2$ (coin flip), two probabilities are $\theta_H$ (head) and $\theta_T$ (tail).
• Here simplex is a line.
• Its dimension: $K - 1 = 2 - 1 = 1$
Bayesian Learning: Multinomial Variable

- Simplex:
- If $K = 3$, three distributions are $\theta_1, \theta_2$ and $\theta_3$.
- Here *simplex is a plane*.
- Its dimension: $K - 1 = 3 - 1 = 2$
Bayesian Learning: Multinomial Variable

• An example: $K = 3$
• The **Dirichlet distribution** over three variables $\theta_1$, $\theta_2$ and $\theta_3$ is confined to a simplex (a bounded linear manifold).
• Points on this **surface** satisfy the following constraints.

\[
\sum_k \theta_k = 1 \\
0 \leq \theta_k \leq 1
\]
Bayesian Learning: Multinomial Variable

- Let’s interpret the Dirichlet parameters $\alpha_0$ and $\alpha_k$.
- The parameter $\alpha_0$ controls the **strength of the distribution** (how peaked it is).
- The parameter $\alpha_k$ control the **location of the peak**.
- Generally, the **higher value of $\alpha$**, the greater “weight” of $x_i$ and the greater amount of the total “mass” is assigned to it.

$$
\text{Dir}(\boldsymbol{\theta} \mid \tilde{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}
$$

$$
\alpha_0 = \sum_{k=1}^{K} \alpha_k
$$
Bayesian Learning: Multinomial Variable

- Let’s look at some examples with $K = 3$.
- Plot of the Dirichlet density when $\alpha = (1, 1, 1)$.
- Since $\alpha_1 = \alpha_2 = \alpha_3 = 1$, the points are **uniformly distributed**.
- They **could be anywhere**!
Bayesian Learning: Multinomial Variable

• Example: $K = 3$
• Here the two horizontal axes are coordinates in the plane of the simplex and the vertical axis corresponds to the value of the density.
• $\{\alpha_1, \alpha_2, \alpha_3\} = \{1, 1, 1\}$
Bayesian Learning: Multinomial Variable

- Example: $K = 3$
- Plot of the Dirichlet density when $\alpha = (2, 2, 2)$
- The **strength of the distribution is centered**.
Bayesian Learning: Multinomial Variable

- Example: $K = 3$
- Plot of the Dirichlet density when $\alpha = (20, 2, 2)$
- Here $\alpha_1$ is the largest.
- Hence determines where the peak will be located.
Bayesian Learning: Multinomial Variable

• Example: $K = 3$
Bayesian Learning: Multinomial Variable

- Example: \( K = 3 \)
- Here the two horizontal axes are coordinates in the plane of the simplex and the vertical axis corresponds to the value of the density.
- \( \{\alpha_1, \alpha_2, \alpha_3\} = \{0.1, 0.1, 0.1\} \)

If \( \alpha_i < 1 \), it can be thought as **anti-weight** that pushes away \( x_i \) toward extremes

Far away from the uniform or peaky!
Bayesian Learning: Multinomial Variable

- Example: $K = 3$
- Here the two horizontal axes are coordinates in the plane of the simplex and the vertical axis corresponds to the value of the density.
- $\{\alpha_1, \alpha_2, \alpha_3\} = \{10, 10, 10\}$

When $\alpha_i$ is high, it attracts $x_i$ toward some central value

Here central is in the sense that all points are concentrated around it, not in the sense that it is symmetrically central.
Bayesian Learning: Multinomial Variable

- To further understand the **how the parameter $\alpha$ controls** the distribution consider the samples from a **5-dimensional symmetric Dirichlet** distribution.
- We generate samples from this Dirichlet distribution.

![Graph of Dirichlet samples](image)
Bayesian Learning: Multinomial Variable

- Here $\alpha = (0.1, \ldots, 0.1)$.
- The distribution is far away from uniform.
- Samples are peaky (towards one of the ends).
- Example: 1\textsuperscript{st} row: either 5, 2\textsuperscript{nd} row: either 1 or 3, 3\textsuperscript{rd} row: 1 or 4

This results in very sparse distributions, with many 0s.
Bayesian Learning: Multinomial Variable

• Samples from a 5-dimensional symmetric Dirichlet distribution.
• Here $\alpha = (1, \ldots, 1)$ or uniform.

This results in more uniform (and dense) distributions.

That’s why samples can be anywhere!
Bayesian Learning: Multinomial Variable

- So far, we have derived the likelihood function and the prior probability function for multinomial distribution:

\[ p(D \mid \hat{\theta}) = \prod_{k=1}^{K} \theta_k^{N_k} \]

\[ p(\hat{\theta} \mid \hat{\alpha}) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \]

Let’s derive the posterior for this distribution:

\[ \text{Dir}(\hat{\theta} \mid \hat{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \]
Bayesian Learning: Multinomial Variable

- Posterior

\[ p(\hat{\theta} | D) \propto p(D | \hat{\theta}) p(\hat{\theta}) \]

\[ p(\hat{\theta} | D) \propto \prod_{k=1}^{K} \theta_k^{N_k} \theta_k^{\alpha_k-1} = \prod_{k=1}^{K} \theta_k^{N_k+\alpha_k-1} \]

\[ = \text{Dir}(\hat{\theta} | \alpha_1 + N_1, \ldots, \alpha_K + N_K) \]

The posterior is obtained by **adding the prior hyperparameters** (pseudo-counts) \( \alpha_k \) to the empirical counts \( N_k \).
Bayesian Learning: Multinomial Variable

- Let’s derive the **MAP estimate (mode)** of this posterior.
- Generally, we set the **derivative of the log likelihood** (wrt $\theta$) to 0 and determine $\theta$.

\[
p(\hat{\theta} \mid D) \propto \prod_{k=1}^{K} \theta_k^{N_k+\alpha_k-1}
\]
Bayesian Learning: Multinomial Variable

• However, we must enforce the constraint that $\Sigma_k \theta_k = 1$.
• We can do this by using a Lagrange multiplier.
• The constrained objective function, or Lagrangian, is given by the log likelihood plus log prior plus the constraint.
• Here, $\lambda$ is the Lagrange multiplier.

$$p(\theta | D) \propto \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1}$$

$$\ell(\theta, \lambda) = \sum_k N_k \log \theta_k + \sum_k (\alpha_k - 1) \log \theta_k + \lambda \left(1 - \sum_k \theta_k\right)$$
Bayesian Learning: Multinomial Variable

• Taking derivatives with respect to $\lambda$ yields the original constraint ($\sum_k \theta_k = 1$):

$$\frac{\partial \ell}{\partial \lambda} = \left( 1 - \sum_k \theta_k \right) = 0$$

Taking derivatives with respect to $\theta_k$ yields

$$\frac{\partial \ell}{\partial \theta_k} = \frac{N'_k}{\theta_k} - \lambda = 0$$

$$N'_k = \lambda \theta_k$$

To simplify the notation we use:

$$N'_k \triangleq N_k + \alpha_k - 1$$

$$\ell(\theta, \lambda) = \sum_k N_k \log \theta_k + \sum_k (\alpha_k - 1) \log \theta_k + \lambda \left( 1 - \sum_k \theta_k \right)$$
Bayesian Learning: Multinomial Variable

- We can solve for $\lambda$ using the **sum-to-one constraint**:

$$N'_k = \lambda \theta_k$$

$$\sum_k N'_k = \lambda \sum_k \theta_k$$

$$N + \alpha_0 - K = \lambda$$

Here $\alpha_0$ is the equivalent to sample size of the prior.
Bayesian Learning: Multinomial Variable

• Thus the **MAP estimate** is given by:

\[
\hat{\theta}_k = \frac{N_k + \alpha_k - 1}{N + \alpha_0 - K}
\]

If we use a uniform prior \(\alpha_k = 1\), we recover the MLE:

\[
\hat{\theta}_k = \frac{N_k}{N}
\]

This is just the **empirical fraction** of times face \(k\) shows up.

\[
N'_k = \lambda \theta_k
\]

\[
N + \alpha_0 - K = \lambda
\]

\[
N'_k \triangleq N_k + \alpha_k - 1
\]
Bayesian Learning: Multinomial Variable

• How do we make a prediction?
• The posterior predictive distribution for a single multinomial trial is given by the following expression:

\[ p(X = j|\mathcal{D}) = \int p(X = j|\theta)p(\theta|\mathcal{D})d\theta \]

It is the expected value of the random variable \( \theta_j \) (the probability of getting j’th side)
Bayesian Learning: Multinomial Variable

- The expected value of the random variable \( \theta_j \) is given by the mean of the Dirichlet distribution.

\[
p(X = j | \mathcal{D}) = \int p(X = j | \theta) p(\theta | \mathcal{D}) d\theta
\]

\[
= \mathbb{E} [\theta_j | \mathcal{D}]
\]

\[
= \frac{\alpha_j + N_j}{\sum_k (\alpha_k + N_k)}
\]

\[
= \frac{\alpha_j + N_j}{\alpha_0 + N}
\]

\[
\mathbb{E} [x_k] = \frac{\alpha_k}{\alpha_0}
\]

\[
\alpha_0 = \sum_{k=1}^{K} \alpha_k = K
\]

\[
\sum_{k=1}^{K} N_k = N
\]
Bayesian Learning: Multinomial Variable

• This expression is very useful.
• Because it enables to avoid extreme prediction (avoids the zero-count problem).
• Recall previously we encountered add-one smoothing when using Bayesian learning to avoid the zero-count (sparse data) problem in the binary scenario (coin model).

\[ p(X = j | D) = \frac{\alpha_j + N_j}{\alpha_0 + N} \]
Bayesian Learning: Multinomial Variable

- In fact, this form of **Bayesian smoothing** is **even more important** in the multinomial case than the binary case.

- Because the likelihood of **data sparsity increases** once we start **partitioning the data into many categories**.

\[ p(X = j | \mathcal{D}) = \frac{\alpha_j + N_j}{\alpha_0 + N} \]
Bayesian Learning: Multinomial Variable

• Let’s look at an application of **Bayesian smoothing** using the Dirichlet-multinomial model.

• **Language modeling.**

• In this application we **predict which words might occur next** in a sequence.
Bayesian Learning: Multinomial Variable

• Here we will take a very simple-minded approach.
• Assume that the i’th word, $X_i \in \{1, \ldots, K\}$, is sampled independently from all the other words using a $\text{Cat}(\theta)$ distribution.
• This is called the bag of words model.
• Given a past sequence of words, how can we predict which one is likely to come next?
Bayesian Learning: Multinomial Variable

• For example, suppose we observe the following sequence (part of a children’s nursery rhyme):

Mary had a little lamb, little lamb, little lamb,
Mary had a little lamb, its fleece as white as snow

Furthermore, suppose our **vocabulary** consists of the following words:

| mary  | lamb  | little | big   | fleece | white  | black  | snow   | rain   | unk   | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|-------|-------|--------|-------|--------|--------|--------|--------|--------|------|-----|---|---|---|---|---|---|---|---|----|----|
Bayesian Learning: Multinomial Variable

• Here *unk* stands for unknown, and represents all other words that do not appear elsewhere on the list.

• To encode each line of the nursery rhyme, we first strip off punctuation, and remove any stop words such as “a”, “as”, “the”, etc.

```
mary lamb little big fleece white black snow rain unk
1  2  3  4  5  6  7  8  9  10
```

Mary had a little lamb, little lamb, little lamb,
Mary had a little lamb, its fleece as white as snow
Bayesian Learning: Multinomial Variable

- We can also perform stemming.
- Stemming reduces words to their base form, such as stripping off the final s in plural words, or the ing from verbs (e.g., running becomes run).
- In this example, no words need stemming.

```
mary lamb little big fleece white black snow rain unk
1  2  3  4  5  6  7  8  9  10
```

Mary had a little lamb, little lamb, little lamb,
Mary had a little lamb, its fleece as white as snow
Finally, we replace each word by its index into the vocabulary to get:

```
1 10 3 2 3 2 3 2
1 10 3 2 10 5 10 6 8
```

Mary had a little lamb, little lamb, little lamb,
Mary had a little lamb, its fleece as white as snow.
Bayesian Learning: Multinomial Variable

- We now **ignore the word order**, and **count how often** each word occurred, resulting in a histogram of word counts:

<table>
<thead>
<tr>
<th>Token Word</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>mary</td>
<td>lamb</td>
<td>little</td>
<td>big</td>
<td>fleece</td>
<td>white</td>
<td>black</td>
<td>snow</td>
<td>rain</td>
<td>unk</td>
</tr>
<tr>
<td>Count</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

1 10 3 2 3 2 3 2
1 10 3 2 10 5 10 6 8

Mary had a little lamb, little lamb, little lamb, Mary had a little lamb, its fleece as white as snow
Bayesian Learning: Multinomial Variable

- Denote the counts by $N_j$.
- If we use a $\text{Dir}(\alpha)$ prior for $\theta$, the posterior predictive is:

\[
p(\tilde{X} = j | D) = E[\theta_j | D] = \frac{\alpha_j + N_j}{\sum_{j'} \alpha_{j'} + N_{j'}} = \frac{1 + N_j}{10 + 17}
\]

If we set $\alpha_j = 1$, we get

\[
p(\tilde{X} = j | D) = \left(\frac{3}{27}, \frac{5}{27}, \frac{5}{27}, \frac{1}{27}, \frac{2}{27}, \frac{2}{27}, \frac{1}{27}, \frac{2}{27}, \frac{1}{27}, \frac{5}{27}\right)
\]
Bayesian Learning: Multinomial Variable

• The **modes** of the predictive distribution are $X = 2$ ("lamb"), $X = 3$ ("little") and $X = 10$ ("unk").

• Note that the words “big”, “black” and “rain” are **predicted to occur with non-zero probability** in the future, even though they have **never been seen before**.

• We will learn more sophisticated language models later.