# Extra material related to Naïve Bayes classifier

Multinomial Distribution and Dirichlet Distribution

# The multivariate normal extends the Gaussian to higher dimensions

- Analogously, the **<u>Multinomial distribution</u>** extends the Binomial distribution to higher dimensions!

- But, how does the dimensionality increase for a discrete variable?

# Number of states

- The Binomial distribution is associated with **BINARY** outcomes.

- The variable can take 2 possible states, $x \in \{0,1\}$

- With a multinomial distribution, we are dealing with a random variable that can take on **MORE** than 2 states!

# Number of states

- With a multinomial distribution, we are dealing with a random variable that can take on **MORE** than 2 states!

- Examples:
  - Canonical example – rolling a 6 sided die
  - Voting with more than 2 political parties

# 1-of-$K$ encoding

- Denote the total number of states as $K$.

- The random variable is represented as a $K$-dimensional vector.

$$\mathbf{x} = \{x_1, x_2, \dots, x_k, \dots, x_K\}$$

- The observed state is then assigned a value of 1: $x_k = 1$
- All other states are set to 0

# For example, if we roll a 4 from a 6 sided die

- The 6 possible states (1 through 6) are encoded as:

$$\mathbf{x} = \{x_1, x_2, x_3, x_4, x_5, x_6\}$$

- If we observe a 4 the elements in the vector take on the values:

$$\mathbf{x} = \{x_1 = 0, x_2 = 0, x_3 = 0, x_4 = 1, x_5 = 0, x_6 = 0\}$$

# For example, if we roll a 4 from a 6 sided die

- The 6 possible states (1 through 6) are encoded as:

$$\mathbf{x} = \{x_1, x_2, x_3, x_4, x_5, x_6\}$$

- If we observe a 4 the elements in the vector take on the values:

$$\mathbf{x} = \{0, 0, 0, 1, 0, 0\}$$

# Define the probability $x_k = 1$ as $\mu_k$

- The distribution of $\mathbf{x}$ is therefore:

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

- Where $\boldsymbol{\mu} = \{\mu_1, \mu_2, \dots, \mu_k, \dots, \mu_K\}$ is the vector of probabilities for each state.

# Now consider observing $N$ independent observations of the random variable

- Similar to the Multivariate normal we can organize the observation of the $K$ states in a matrix, $\mathbf{X}$.

- The $n$-th observation of the $k$-th state, $x_{n,k}$, will be 0 or 1.

# The likelihood of $\mathbf{X}$ given $\boldsymbol{\mu}$ can be factored into the product of $N$ separate likelihoods

$$p(\mathbf{X}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \{p(\mathbf{x}_{n,:}^{T}|\boldsymbol{\mu})\} = \prod_{n=1}^{N} \left\{ \prod_{k=1}^{K} \mu_k^{x_{n,k}} \right\}$$

# The likelihood can be rearranged as

$$\prod_{n=1}^{N}\left\{\prod_{k=1}^{K}\mu_k^{x_{n,k}}\right\} = \prod_{k=1}^{K}\mu_k^{x_{1,k}} \times \mu_k^{x_{2,k}} \times \cdots \times \mu_k^{x_{n,k}} \times \cdots \times \mu_k^{x_{N,k}}$$

# The likelihood can be rearranged as

$$\prod_{n=1}^{N}\left\{\prod_{k=1}^{K}\mu_k^{x_{n,k}}\right\} = \prod_{k=1}^{K}\underbrace{\mu_k^{x_{1,k}} \times \mu_k^{x_{2,k}} \times \cdots \times \mu_k^{x_{n,k}} \times \cdots \times \mu_k^{x_{N,k}}}$$

$$\prod_{n=1}^{N}\left\{\prod_{k=1}^{K}\mu_k^{x_{n,k}}\right\} = \prod_{k=1}^{K}\mu_k^{\left(\sum_{n=1}^{N} x_{n,k}\right)}$$

# Sufficient statistics…are just counting!

- Define the number of times $x_k = 1$ as:

$$m_k = \sum_{n=1}^{N} x_{n,k}$$

# The likelihood of the observations given the state probabilities is therefore:

$$p(\mathbf{X}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \left\{ p\left(\mathbf{x}_{n,:}^{T}|\boldsymbol{\mu}\right) \right\} = \prod_{k=1}^{K} \mu_k^{m_k}$$

## What are we still missing…remember how we went from the Bernoulli to the Binomial for the binary outcome case?

- Just as we saw with the binary outcome situation, there are multiple potential sequences for observing exactly $m_k$ counts out of $N$ trials.

- Therefore, we need to account for the number of ways of partitioning $N$ objects into $K$ groups of size $m_1, m_2, \dots, m_K$.

# The multinomial distribution

$$p(m_1, m_2, \ldots, m_K | \boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \cdots m_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

# Without deriving the MLE on $\boldsymbol{\mu}$ can you guess what it is?

HINT: The basic definition of probability…

# The MLE on the vector probabilities per state

$$\hat{\boldsymbol{\mu}} = \{\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_K\} = \left\{\frac{m_1}{N}, \frac{m_2}{N}, \dots, \frac{m_K}{N}\right\}$$

# Bayesian formulation – prior specification

- We saw in the Binary case, that the conjugate prior for the Binomial likelihood is the Beta distribution.
- Since the Multinomial is a multivariate generalization of the Binomial, we can expect that the corresponding conjugate prior is a multivariate generalization of the Beta…

# Bayesian formulation – prior specification

- We saw in the Binary case, that the conjugate prior for the Binomial likelihood is the Beta distribution.
- Since the Multinomial is a multivariate generalization of the Binomial, we can expect that the corresponding conjugate prior is a multivariate generalization of the Beta…

**Dirichlet distribution**

# The Dirichlet distribution

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

# The Dirichlet distribution…is confined to a <u>simplex</u>

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

The simplex results from the summation constraint on the state probabilities: $\sum_k \mu_k = 1$