The Dirichlet-Discrete Model

Readings - Bishop: Section 2.2 and Appendix E
The first six pages of the optional reading by Frigyik, Kapila, and Gupta are also recommended

- The third class discussed the Beta-Bernoulli Model

- This class will generalize that model from binary random variables to variables taking values in a finite set (often called "categorical" or "discrete" variables)

- For example, the set of words in a vocabulary

- For simplicity, we will denote this set as $\{1, 2, \ldots, V\}$ where V is at least 2 and known in advance (the case where V is not known in advance is a topic for a more advanced class)

Discrete Random Variables

- The Beta-Bernoulli model had a single parameter, μ

- Now μ becomes a vector with V components:
  $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_V]$ with $\mu_i \geq 0$ and $\sum \mu_i = 1$

- The set of all legal $\boldsymbol{\mu}$ is denoted Δ^V

-

Discrete Random Variables

- The Beta-Bernoulli model had a single parameter, μ
- Now μ becomes a vector with V components:
  $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_V]$ with $\mu_i \geq 0$ and $\sum \mu_i = 1$
- The set of all legal $\boldsymbol{\mu}$ is denoted Δ^V
- The value of a discrete random variable can be represented by a "one-hot" vector with V components:
  $[0, 0, \ldots, 0, 1, 0, \ldots, 0]$, where the position of the 1 indicates the value of the variable

Discrete Random Variables

- The Beta-Bernoulli model had a single parameter, μ

- Now μ becomes a vector with V components:
  $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_V]$ with $\mu_i \geq 0$ and $\sum \mu_i = 1$

- The set of all legal $\boldsymbol{\mu}$ is denoted Δ^V

- The value of a discrete random variable can be represented by a "one-hot" vector with V components:
  [0, 0, …, 0, 1, 0, …, 0], where the position of the 1 indicates the value of the variable

- If X_i is a discrete random variable, we can express its probability distribution as DiscretePMF$(X = w) = \prod_i \mu_i^{X_{wi}}$
  where $X_{wi} = 1$ only when $X_i = w$

The Likelihood Function

- Suppose we observe N words from a vocabulary of size V and denote the random variables associated with these observations by $X_1, X_2, \ldots, X_N$

-

The Likelihood Function

- Suppose we observe N words from a vocabulary of size V and denote the random variables associated with these observations by $X_1, X_2, \ldots, X_N$

- Although these words will not usually be independent, we can make a simplifying assumption (called "bag-of-words") that the X_i are i.i.d. (independent and identically distributed)

The Likelihood Function

- Suppose we observe N words from a vocabulary of size V and denote the random variables associated with these observations by $X_1, X_2, \ldots, X_N$

- Although these words will not usually be independent, we can make a simplifying assumption (called "bag-of-words") that the X_i are i.i.d. (independent and identically distributed)

- This gives the likelihood function

$$P(X_1, X_2, \ldots, X_N | \boldsymbol{\mu}) = \prod_n \prod_i \mu_i^{X_{ni}} = \prod_i \mu_i^{m_i}$$

where $m_i = \sum_n X_{ni}$ is a count of the number of times word i appears in the dataset

A Maximum Likelihood Estimate for μ

- Since ln(x), the natural logarithm of x, is an increasing function of x, we can maximize the log-likelihood instead of maximizing the likelihood directly. This simplifies the math and helps prevent numerical problems.
-

A Maximum Likelihood Estimate for $\boldsymbol{\mu}$

- Since ln(x), the natural logarithm of x, is an increasing function of x, we can maximize the log-likelihood instead of maximizing the likelihood directly. This simplifies the math and helps prevent numerical problems.

- $\boldsymbol{\mu}^{ML} = \arg\max \sum_i m_i \ln \mu_i$

  where the maximum is over all $\boldsymbol{\mu}$ in Δ^V

A Maximum Likelihood Estimate for $\boldsymbol{\mu}$

- Since ln(x), the natural logarithm of x, is an increasing function of x, we can maximize the log-likelihood instead of maximizing the likelihood directly. This simplifies the math and helps prevent numerical problems.

- $\boldsymbol{\mu}^{ML} = \arg\max \sum_i m_i \ln \mu_i$

  where the maximum is over all $\boldsymbol{\mu}$ in Δ^V

- Since Δ^V is the V-1 dimensional subspace of legal V dimensional probability vectors, this is a constrained optimization problem and we can use *Lagrange multipliers* to find the $\boldsymbol{\mu}$ that gives the maximum of the likelihood function

Lagrange Multipliers

- Suppose that we want to maximize $f(\mathbf{x})$ subject to a constraint $g(\mathbf{x}) = 0$, where $\mathbf{x}$ is a D-dimensional vector
-
-
-
-

Lagrange Multipliers

- Suppose that we want to maximize $f(\mathbf{x})$ subject to a constraint $g(\mathbf{x}) = 0$, where $\mathbf{x}$ is a D-dimensional vector

- $g(\mathbf{x}) = 0$ defines a D-1 dimensional surface and the gradient $\nabla g(\mathbf{x})$ is perpendicular to this surface

Lagrange Multipliers

- Suppose that we want to maximize $f(\mathbf{x})$ subject to a constraint $g(\mathbf{x}) = 0$, where $\mathbf{x}$ is a D-dimensional vector

- $g(\mathbf{x}) = 0$ defines a D-1 dimensional surface and the gradient $\nabla g(\mathbf{x})$ is perpendicular to this surface

- At the point where $f(\mathbf{x})$ reaches its maximum, $\nabla f(\mathbf{x})$ must also be perpendicular to the surface - if the gradient had any component along the surface we could move along the surface in this direction to get to a larger value for $f(\mathbf{x})$

-

-

Lagrange Multipliers

- Suppose that we want to maximize $f(\mathbf{x})$ subject to a constraint $g(\mathbf{x}) = 0$, where $\mathbf{x}$ is a D-dimensional vector

- $g(\mathbf{x}) = 0$ defines a D-1 dimensional surface and the gradient $\nabla g(\mathbf{x})$ is perpendicular to this surface

- At the point where $f(\mathbf{x})$ reaches its maximum, $\nabla f(\mathbf{x})$ must also be perpendicular to the surface - if the gradient had any component along the surface we could move along the surface in this direction to get to a larger value for $f(\mathbf{x})$

- Since these two gradient vectors point in the same (or exact opposite) direction, we can write $\nabla f(\mathbf{x}) + \lambda \nabla g(\mathbf{x}) = 0$

-

Lagrange Multipliers

- Suppose that we want to maximize $f(\mathbf{x})$ subject to a constraint $g(\mathbf{x}) = 0$, where $\mathbf{x}$ is a D-dimensional vector
- $g(\mathbf{x}) = 0$ defines a D-1 dimensional surface and the gradient $\nabla g(\mathbf{x})$ is perpendicular to this surface
- At the point where $f(\mathbf{x})$ reaches its maximum, $\nabla f(\mathbf{x})$ must also be perpendicular to the surface - if the gradient had any component along the surface we could move along the surface in this direction to get to a larger value for $f(\mathbf{x})$
- Since these two gradient vectors point in the same (or exact opposite) direction, we can write $\nabla f(\mathbf{x}) + \lambda \nabla g(\mathbf{x}) = 0$
- This motivates the definition of the Lagrangian:

$$\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})$$

Lagrange Multipliers

- Suppose that we want to maximize $f(\mathbf{x})$ subject to a constraint $g(\mathbf{x}) = 0$, where $\mathbf{x}$ is a D-dimensional vector

- $g(\mathbf{x}) = 0$ defines a D-1 dimensional surface and the gradient $\nabla g(\mathbf{x})$ is perpendicular to this surface

- At the point where $f(\mathbf{x})$ reaches its maximum, $\nabla f(\mathbf{x})$ must also be perpendicular to the surface - if the gradient had any component along the surface we could move along the surface in this direction to get to a larger value for $f(\mathbf{x})$

- Since these two gradient vectors point in the same (or exact opposite) direction, we can write $\nabla f(\mathbf{x}) + \lambda \nabla g(\mathbf{x}) = 0$

- This motivates the definition of the Lagrangian:
$$\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})$$

- To maximize $f(\mathbf{x})$ subject to $g(\mathbf{x}) = 0$, take partial derivatives of $\mathcal{L}(\mathbf{x}, \lambda)$ with respect to λ and the components of $\mathbf{x}$ and set these derivatives to zero

Using Lagrange Multipliers to Find a Maximum Likelihood Estimate for $\boldsymbol{\mu}$

- First write the Lagrangian:
  $\mathcal{L}(\mu_1, \mu_2, \ldots, \mu_V, \lambda) = \sum_i m_i \ln \mu_i + \lambda(1 - \sum_i \mu_i)$
-
-
-

Using Lagrange Multipliers to Find a Maximum Likelihood Estimate for $\boldsymbol{\mu}$

- First write the Lagrangian:
  $\mathcal{L}(\mu_1, \mu_2, \ldots, \mu_V, \lambda) = \sum_i m_i \ln \mu_i + \lambda(1 - \sum_i \mu_i)$

- The partial derivative with respect to λ just gives back the constraint $\sum_i \mu_i = 1$

-

-

Using Lagrange Multipliers to Find a Maximum Likelihood Estimate for $\boldsymbol{\mu}$

- First write the Lagrangian:
  $\mathcal{L}(\mu_1, \mu_2, \ldots, \mu_V, \lambda) = \sum_i m_i \ln \mu_i + \lambda(1 - \sum_i \mu_i)$

- The partial derivative with respect to λ just gives back the constraint $\sum_i \mu_i = 1$

- The partial derivatives with respect to μ_i give the equations
  $m_i/\mu_i - \lambda = 0$, or $\mu_i = m_i/\lambda$

-

Using Lagrange Multipliers to Find a Maximum Likelihood Estimate for $\boldsymbol{\mu}$

- First write the Lagrangian:
  $\mathcal{L}(\mu_1, \mu_2, \ldots, \mu_V, \lambda) = \sum_i m_i \ln \mu_i + \lambda(1 - \sum_i \mu_i)$

- The partial derivative with respect to λ just gives back the constraint $\sum_i \mu_i = 1$

- The partial derivatives with respect to μ_i give the equations
  $m_i/\mu_i - \lambda = 0$, or $\mu_i = m_i/\lambda$

- Plugging these values of μ_i into the constraint gives
  $\lambda = \sum_i m_i = N$

Using Lagrange Multipliers to Find a Maximum Likelihood Estimate for $\boldsymbol{\mu}$

- First write the Lagrangian: $\mathcal{L}(\mu_1, \mu_2, \ldots, \mu_V, \lambda) = \sum_i m_i \ln \mu_i + \lambda(1 - \sum_i \mu_i)$
- The partial derivative with respect to λ just gives back the constraint $\sum_i \mu_i = 1$
- The partial derivatives with respect to μ_i give the equations $m_i/\mu_i - \lambda = 0$, or $\mu_i = m_i/\lambda$
- Plugging these values of μ_i into the constraint gives $\lambda = \sum_i m_i = N$
- Putting this all together gives $\boldsymbol{\mu}^{ML} = [m_1/N, m_2/N, \ldots, m_V/N]$ which has all its components in the interval [0,1] as desired

Sufficient Statistics and the Multinomial Distribution

- This means that (under the bag-of-words assumption) all we need to know about the data is contained in the quantities m_i so the m_i are called *sufficient statistics* for $\boldsymbol{\mu}^{ML}$

-

Sufficient Statistics and the Multinomial Distribution

- This means that (under the bag-of-words assumption) all we need to know about the data is contained in the quantities m_i so the m_i are called *sufficient statistics* for $\boldsymbol{\mu}^{ML}$

- The distribution of the m_i values, conditioned on $\boldsymbol{\mu}$ and N is *multinomial*:

$$\mathrm{Mult}(m_1, m_2, \ldots, m_V \mid \boldsymbol{\mu}, N) = C(N; m_1, m_2, \ldots, m_V) \prod_i \mu_i^{m_i}$$

where $C(N; m_1, m_2, \ldots, m_V) = N! / (m_1!\, m_2! \ldots m_V!)$
are the *multinomial coefficients* found in the expansion of
$(x_1 + x_2 + \ldots + x_V)^N$

Sufficient Statistics and the Multinomial Distribution

- This means that (under the bag-of-words assumption) all we need to know about the data is contained in the quantities m_i so the m_i are called *sufficient statistics* for $\boldsymbol{\mu}^{ML}$

- The distribution of the m_i values, conditioned on $\boldsymbol{\mu}$ and N is *multinomial*:

$$\mathrm{Mult}(m_1, m_2, \ldots, m_V \mid \boldsymbol{\mu}, N) = C(N; m_1, m_2, \ldots, m_V) \prod_i \mu_i^{m_i}$$

where $C(N; m_1, m_2, \ldots, m_V) = N! \,/\, (m_1!\, m_2! \ldots m_V!)$
are the *multinomial coefficients* found in the expansion of
$(x_1 + x_2 + \ldots + x_V)^N$

- Continuing the analogy to the Beta-Bernoulli model, we can generalize the multinomial distribution to the *Dirichlet* distribution, again replacing the factorials with gamma functions

The Dirichlet Distribution

- The Dirichlet distribution is a conjugate prior for the parameters of the multinomial distribution
-
-

The Dirichlet Distribution

- The Dirichlet distribution is a conjugate prior for the parameters of the multinomial distribution

- The probability density function (PDF) of the Dirichlet distribution is given by

  $$\mathrm{DirPDF}(\boldsymbol{\mu} \mid a_1, a_2, \ldots, a_V) = c(\mathbf{a}) \prod_i \mu_i^{a_i - 1}$$

  where the normalizing factor $c(\mathbf{a}) = \Gamma(\sum_i a_i)/\prod_i \Gamma(a_i)$ generalizes the multinomial coefficient by replacing the factorials with gamma functions

-

The Dirichlet Distribution

- The Dirichlet distribution is a conjugate prior for the parameters of the multinomial distribution

- The probability density function (PDF) of the Dirichlet distribution is given by

  $$\mathrm{DirPDF}(\boldsymbol{\mu} \mid a_1, a_2, \ldots, a_V) = c(\mathbf{a}) \prod_i \mu_i^{a_i-1}$$

  where the normalizing factor $c(\mathbf{a}) = \Gamma(\sum_i a_i)/\prod_i \Gamma(a_i)$ generalizes the multinomial coefficient by replacing the factorials with gamma functions

- When V=2, this is the Beta distribution

The Dirichlet Distribution

- The Dirichlet distribution is a conjugate prior for the parameters of the multinomial distribution

- The probability density function (PDF) of the Dirichlet distribution is given by

  $$\mathrm{DirPDF}(\boldsymbol{\mu} \mid a_1, a_2, \ldots, a_V) = c(\mathbf{a}) \prod_i \mu_i^{a_i-1}$$

  where the normalizing factor $c(\mathbf{a}) = \Gamma(\sum_i a_i)/\prod_i \Gamma(a_i)$ generalizes the multinomial coefficient by replacing the factorials with gamma functions

- When V=2, this is the Beta distribution

- The Dirichlet distribution is defined on the probability simplex given by the constraints $\sum_i \mu_i = 1$ and $\mu_i \geq 0$

Distributions Derived from the Dirichlet-Discrete Model

- The Dirichlet-Discrete joint distribution defines a complete model:

  $P(X_1, X_2, \ldots, X_N, \boldsymbol{\mu}) =$
  $[\prod_n \text{DiscretePMF}(X_n \mid \boldsymbol{\mu})]\ \text{DirPDF}(\boldsymbol{\mu} \mid a_1, a_2, \ldots, a_V)$
  where the first factor is the likelihood and the second is the prior

-

Distributions Derived from the Dirichlet-Discrete Model

- The Dirichlet-Discrete joint distribution defines a complete model:

  $P(X_1, X_2, \ldots, X_N, \boldsymbol{\mu}) =$
  $\quad [\prod_n \text{DiscretePMF}(X_n \mid \boldsymbol{\mu})]\ \text{DirPDF}(\boldsymbol{\mu} \mid a_1, a_2, \ldots, a_V)$
  where the first factor is the likelihood and the second is the prior

- Several distributions can be derived from this:
  — Evidence: $P(X_1, X_2, \ldots, X_N) = \int_{\Delta V} P(X_1, X_2, \ldots, X_N, \boldsymbol{\mu})\, d\boldsymbol{\mu}$

  —

  —

Distributions Derived from the Dirichlet-Discrete Model

- The Dirichlet-Discrete joint distribution defines a complete model:

  $P(X_1, X_2, \ldots, X_N, \boldsymbol{\mu}) =$
  $\quad [\prod_n \text{DiscretePMF}(X_n \mid \boldsymbol{\mu})]\ \text{DirPDF}(\boldsymbol{\mu} \mid a_1, a_2, \ldots, a_V)$
  where the first factor is the likelihood and the second is the prior

- Several distributions can be derived from this:
  — Evidence: $P(X_1, X_2, \ldots, X_N) = \int_{\Delta V} P(X_1, X_2, \ldots, X_N, \boldsymbol{\mu})\, d\boldsymbol{\mu}$

  — Posterior: $P(\boldsymbol{\mu} \mid X_1, X_2, \ldots, X_N)$, obtained by dividing the joint distribution by the evidence (Bayes rule)

  —

Distributions Derived from the Dirichlet-Discrete Model

- The Dirichlet-Discrete joint distribution defines a complete model:

  $P(X_1, X_2, \ldots, X_N, \boldsymbol{\mu}) =$
  $\quad [\prod_n \text{DiscretePMF}(X_n \mid \boldsymbol{\mu})] \; \text{DirPDF}(\boldsymbol{\mu} \mid a_1, a_2, \ldots, a_V)$
  where the first factor is the likelihood and the second is the prior

- Several distributions can be derived from this:
  — Evidence: $P(X_1, X_2, \ldots, X_N) = \int_{\Delta V} P(X_1, X_2, \ldots, X_N, \boldsymbol{\mu}) \, d\boldsymbol{\mu}$

  — Posterior: $P(\boldsymbol{\mu} \mid X_1, X_2, \ldots, X_N)$, obtained by dividing the joint distribution by the evidence (Bayes rule)

  — Predictive Posterior: $P(X_N \mid X_1, X_2, \ldots, X_{N-1}) =$
  $\int_{\Delta V} P(X_N \mid \boldsymbol{\mu}) \, P(\boldsymbol{\mu} \mid X_1, X_2, \ldots, X_{N-1}) \, d\boldsymbol{\mu}$

Deriving the MAP Estimate from the Posterior

- Suppose we have some prior knowledge of $\boldsymbol{\mu}$, represented as a prior distribution, and we want to combine this with new data X_1, X_2, …, X_N, to obtain a posterior distribution for $\boldsymbol{\mu}$ and use this to get a *maximum a posteriori* (MAP) estimate for $\boldsymbol{\mu}$

-

Deriving the MAP Estimate from the Posterior

- Suppose we have some prior knowledge of $\boldsymbol{\mu}$, represented as a prior distribution, and we want to combine this with new data $X_1, X_2, \ldots, X_N$, to obtain a posterior distribution for $\boldsymbol{\mu}$ and use this to get a *maximum a posteriori* (MAP) estimate for $\boldsymbol{\mu}$

- The evidence, $P(X_1, X_2, \ldots, X_N)$, and the normalizing factor, $c(\mathbf{a})$, don't depend on $\boldsymbol{\mu}$, so they don't affect the MAP estimate:

$$P(\boldsymbol{\mu} \mid X_1, X_2, \ldots, X_N) \sim$$
$$[\textstyle\prod_n \text{DiscretePMF}(X_n \mid \boldsymbol{\mu})]\ \text{DirPDF}(\boldsymbol{\mu} \mid a_1, a_2, \ldots, a_V) \sim$$

where $\sim$ means that factors not involving $\boldsymbol{\mu}$ have been omitted

Deriving the MAP Estimate from the Posterior

- Suppose we have some prior knowledge of $\boldsymbol{\mu}$, represented as a prior distribution, and we want to combine this with new data $X_1, X_2, \ldots, X_N$, to obtain a posterior distribution for $\boldsymbol{\mu}$ and use this to get a *maximum a posteriori* (MAP) estimate for $\boldsymbol{\mu}$

- The evidence, $P(X_1, X_2, \ldots, X_N)$, and the normalizing factor, $c(\mathbf{a})$, don't depend on $\boldsymbol{\mu}$, so they don't affect the MAP estimate:

$$P(\boldsymbol{\mu} \mid X_1, X_2, \ldots, X_N) \sim$$
$$[\textstyle\prod_n \text{DiscretePMF}(X_n \mid \boldsymbol{\mu})]\ \text{DirPDF}(\boldsymbol{\mu} \mid a_1, a_2, \ldots, a_V) \sim$$
$$[\textstyle\prod_n \prod_i \mu_i^{X_{ni}}] \cdot \prod_i \mu_i^{a_i-1} =$$

where $\sim$ means that factors not involving $\boldsymbol{\mu}$ have been omitted

Deriving the MAP Estimate from the Posterior

- Suppose we have some prior knowledge of $\boldsymbol{\mu}$, represented as a prior distribution, and we want to combine this with new data $X_1, X_2, \ldots, X_N$, to obtain a posterior distribution for $\boldsymbol{\mu}$ and use this to get a *maximum a posteriori* (MAP) estimate for $\boldsymbol{\mu}$

- The evidence, $P(X_1, X_2, \ldots, X_N)$, and the normalizing factor, $c(\mathbf{a})$, don't depend on $\boldsymbol{\mu}$, so they don't affect the MAP estimate:

$$P(\boldsymbol{\mu} \mid X_1, X_2, \ldots, X_N) \sim$$
$$\left[\prod_n \text{DiscretePMF}(X_n \mid \boldsymbol{\mu})\right] \text{DirPDF}(\boldsymbol{\mu} \mid a_1, a_2, \ldots, a_V) \sim$$
$$\left[\prod_n \prod_i \mu_i^{X_{ni}} \right] \cdot \prod_i \mu_i^{a_i-1} =$$
$$\prod_i \mu_i^{m_i+a_i-1}$$

where ~ means that factors not involving $\boldsymbol{\mu}$ have been omitted

Deriving the MAP Estimate from the Posterior

- $\boldsymbol{\mu}^{MAP} = \arg\max P(\boldsymbol{\mu} \mid X_1, X_2, \ldots, X_N)$
  $= \arg\max \prod_i \mu_i^{m_i+a_i-1}$
  $= \arg\max \sum_i (m_i+a_i-1) \ln \mu_i$

-

Deriving the MAP Estimate from the Posterior

- $\boldsymbol{\mu}^{\mathrm{MAP}} = \arg\max P(\boldsymbol{\mu} \mid X_1, X_2, \ldots, X_N)$
  $= \arg\max \prod_i \mu_i^{m_i+a_i-1}$
  $= \arg\max \sum_i (m_i+a_i-1) \ln \mu_i$

- Note that the a_i need to be at least 1 to ensure that the coefficients (m_i+a_i-1) are non-negative

Deriving the MAP Estimate from the Posterior

- $\boldsymbol{\mu}^{MAP} = \arg\max P(\boldsymbol{\mu} \mid X_1, X_2, \ldots, X_N)$
  $= \arg\max \prod_i \mu_i^{m_i+a_i-1}$
  $= \arg\max \sum_i (m_i+a_i-1) \ln \mu_i$

- Note that the a_i need to be at least 1 to ensure that the coefficients (m_i+a_i-1) are non-negative

- Given that

  $\boldsymbol{\mu}^{MAP} = \arg\max \sum_i (m_i+a_i-1) \ln \mu_i$

  we can again use Lagrange multipliers to obtain

  $\boldsymbol{\mu}^{MAP} = [(m_1+a_1-1)/(N+\sum_i(a_i-1), \ldots, (m_V+a_V-1)/(N+\sum_i(a_i-1)]$