CSE 4/574
Final 2008 Fall

Machine Learning TA's
Department of Computer Science
University at Buffalo
Amherst, NY 14260
changsu,pdas3@buffalo.edu

1 Modeling words using Multinomial Dirichlet

Imagine you are a Google engineer who likes to play with words. Your company has crawled and collected M documents with a total of N words and a vocabulary size of V. Note that N can be greater than V. You are interested in modeling the probability distrbution of words and predicting probability masses of words in new documents.

1. (6 marks) Assume the each word w_j, $j = 1, ..., V$ has probability of occurence β_j. Thus is w_j is observed $w^{(j)}$ times in the corpus, then $p(w_j|\beta_j) = \beta_j^{w^{(j)}}$. Under this model, write down the log likelihood function of your dataset and find the Maximum Likelihood (ML) estimate of β_i. (Hint: You will need to add the term $\lambda(\sum_{j=1}^{V} \beta_j - 1)$ to your loglikelihood function and use the technique for optimization using lagrange multiplier, which here is λ)

2. (5 marks) Now assume that you are given a prior over the parameters $\boldsymbol{\beta}$ collected from previous crawls. The prior probability density function is given by
$$p(\boldsymbol{\beta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{j=1}^{V} \alpha_j)}{\prod_{j=1}^{V} \Gamma(\alpha_j)} \prod_{j=1}^{V} \beta_j^{\alpha_j - 1}$$
Find the posterior distribution of $\boldsymbol{\beta}$ and explain whether the posterior distribution has any similarity to the prior distribution.

3. (5 marks) Now you receive a new document and you try to predict the probability mass of the first word in the document which happens to be word w_i in the vocabulary. Derive an expression of the predictive probability mass associated with word w_i in the new document.

4. (4 marks) What is the expectation and variance for the posterior distribution for β. (Hint: The variance of the distribution of $\beta_{Posterior}$ is $Var[x] = E[x^2] - (E[x])^2$. Also for a Gamma ($\Gamma$) function, $\Gamma(x+1) = x\Gamma(x)$)

Solution

1.
$$p(\mathbf{w}|\boldsymbol{\beta}) = \prod_{j=1}^{V} \beta_j^{w^{(j)}}$$
$$\log L = \sum_{j=1}^{V} w^{(j)} \log \beta_j$$
$$\Rightarrow for\, \beta_{i_{ML}},\ \frac{\partial \log L}{\partial \beta_i} + \frac{\partial \lambda(\sum_{j=1}^{V} \beta_j - 1)}{\partial \beta_i} = 0$$
$$\Rightarrow \lambda = -N\ since\ \sum_{j=1}^{V} \beta_j = 1$$
$$\Rightarrow \beta_{i_{ML}} = \frac{w^{(i)}}{N}$$

2.

$$p(\boldsymbol{\beta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{j=1}^{V})\alpha_j}{\prod_{j=1}^{V}\Gamma(\alpha_j)} \prod_{j=1}^{V} \beta_j^{\alpha_j - 1}$$

$$p(\boldsymbol{\beta}|\boldsymbol{\alpha}, \mathbf{w}) = \frac{p(\mathbf{w}|\boldsymbol{\beta})p(\boldsymbol{\beta}|\boldsymbol{\alpha})}{\int p(\mathbf{w}|\boldsymbol{\beta})p(\boldsymbol{\beta}|\boldsymbol{\alpha})}$$

$$\frac{C(\boldsymbol{\alpha}) \prod_{j=1}^{V} \beta_j^{w^{(j)} + \alpha_j - 1}}{\frac{C(\boldsymbol{\alpha})}{C(\boldsymbol{\alpha} + \mathbf{w}^{(.)})}}$$

$$where\, C(\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{j=1}^{V} \alpha_j)}{\prod_{j=1}^{V}\Gamma(\alpha_j)}$$

$$and\ C(\boldsymbol{\alpha} + \mathbf{w}^{(.)}) = \frac{\sum_{j=1}^{V}\Gamma(\alpha_j + w^{(.)})}{\prod_{j=1}^{V}\Gamma(\alpha_j + w^{(.)})}$$

$$= Dir(\boldsymbol{\beta}|\boldsymbol{\alpha} + \mathbf{w}^{(.)})$$

3. To calculate the predictive prob. mass for w_i in the new document

$$p(w_i|\mathbf{w}, \boldsymbol{\alpha}) = \int d\boldsymbol{\beta} p(\mathbf{w_i}|\boldsymbol{\beta}) p(\boldsymbol{\beta}|\mathbf{w}, \boldsymbol{\alpha})$$

$$= \int d\boldsymbol{\beta} \beta_i C(\boldsymbol{\alpha} + \mathbf{w}^{(.)}) \prod_{j=1}^{V} \Gamma(\alpha_j + w^{(.)})$$

$$= C(\boldsymbol{\alpha} + \mathbf{w}^{(.)}) \int d\boldsymbol{\beta}\, \beta_i^{w^{(i)} + \alpha_i + 1 - 1} \prod_{j=1, j\neq i}^{V} \beta_j^{w^{(j)} + \alpha_j - 1}$$

$$= \frac{C(\alpha_1 + w^{(1)}, ..., \alpha_i + w^{(i)}, ..., \alpha_v + w^{(V)})}{C(\alpha_1 + w^{(1)}, ..., \alpha_i + w^{(i)} + 1, ..., \alpha_v + w^{(V)})}$$

$$= \frac{\alpha_i + w^{(i)}}{\sum_{j=1}^{V}(\alpha_j + w^{(j)})}$$

4. For the variance of the distribution of $\beta_{posterior}$, we have

$$E[\beta_i] = \frac{\alpha_i + w^{(i)}}{\sum_{j=1}^{V}(\alpha_j + w^{(j)})} = \frac{d_i}{\sum_{j=1}^{V} d_j}\ say$$

$$E[\beta_i^2] = \frac{d_i(d_i + 1)}{\sum_{j=1}^{V} d_j(d_j + 1)}$$

$$var[\beta_i] = E[\beta_i^2] - (E[\beta_i])^2$$

2 Neural Networks, Kernel Methods

Imagine you are being asked to participate in a competition to classify a set of user choices, represented as a set of I dimensional features, into one of K output classes of movie categories. For this problem, you plan to use a Neural Network for classification with the most fancy hidden neuron activation function and also model the error function as $G(\mathbf{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_k^{(n)} ln\, y_k^{(n)}$, where the training set is $D = \{\mathbf{x}^{(n)}, t^{(n)}\}$ for $n = 1, ..., N$ and $t^{(n)}$ is a k-ary classification $\{0, ..., k-1\}$ i.e. 1-of-K vector. You also choose the number of hidden neurons to be J. The equations for the two layers of the network are:

- Input to Hidden: $a_j = \sum_{i=1}^{I} w_{ij} x_i \qquad h_j = f^{(1)}(a_j)$
- Hidden to Output: $a_k = \sum_{j=1}^{J} w_{jk} h_j \qquad y_k = f^{(2)}(a_k)$
- The functions $f^{(1)}(.)$ and $f^{(2)}(.)$ are:

$$f^{(1)}(r) = \frac{1}{1 + \exp\{-\frac{e^r - e^{-r}}{e^r + e^{-r}}\}}$$

and

$$f^{(2)}(a_k) = \frac{e^{c_k . a_k}}{\sum_{k'=1}^{K} e^{c_{k'} . a_{k'}}}$$

1. (6 marks) Find the derivatives of the error function w.r.t the hidden to output weights w_{jk}

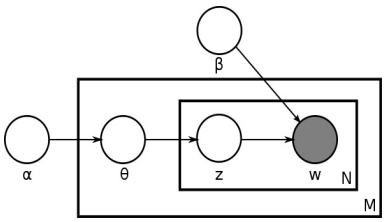

Figure 1: Topic Model

2. (6 marks) Find the derivatives of the error function w.r.t the input to hidden weights w_{ij} .
3. (4 marks) What is the necessary and sufficient condition for a Kernel $k(\mathbf{x}, \mathbf{x}')$ to be valid where $\mathbf{x}$ and $\mathbf{x}'$ are any two vectors of reals? Use this fact to show that for valid Kernels $k_1(\mathbf{x}, \mathbf{x}')$ and $k_2(\mathbf{x}, \mathbf{x}')$, $k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}')$ is a valid Kernel
4. (4 marks) Assuming $poly(k_1(\mathbf{x}, \mathbf{x}'))$ is a valid kernel, where the coefficients of the polynomial are positive, show that $exp(k_1(\mathbf{x}, \mathbf{x}'))$ is a valid Kernel. [HINT: $e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$]

Solution

1. Derivative of Error function w.r.t Hidden to Output weights w_{jk}:

$$\frac{\partial G(\mathbf{w})}{\partial w_{jm}} = \frac{\partial G(\mathbf{w})}{\partial a_m} . \frac{\partial a_m}{\partial w_{jm}} \; for \, m \in 1, ..., K$$

(taking derivative of $\log y_m^{(n)}$ w.r.t a_m instead of $y_m^{(n)}$ w.r.t a_m)

$$= -\sum_{n=1}^{N} \sum_{k=1}^{K} t_k^{(n)} [c_k \delta(m, k) - c_{k''} y_{k''} \delta(m, k'')] h_j$$

$$= -\sum_{n=1}^{N} [t_m^{(n)} . c_m - c_m y_m^{(n)} [\sum_{k=1}^{K} t_k^{(n)}]] h_j \qquad where \qquad \sum_{k=1}^{K} t_k^{(n)} = 1$$

$$\therefore \frac{\partial G(\mathbf{w})}{\partial w_{jk}} = -\sum_{n=1}^{N} c_k [t_k^{(n)} - y_k^{(n)}] h_j$$

2. Derivative of Error function w.r.t Input to Hidden weights w_{ij}:

$$\frac{\partial G(\mathbf{w})}{\partial w_{ij}} = \frac{\partial G(\mathbf{w})}{\partial a_k} . \frac{\partial a_k}{\partial h_j} . \frac{\partial h_j}{\partial a_j} . \frac{\partial a_j}{\partial w_{ij}}$$

$$= -\sum_{n=1}^{N} \sum_{k=1}^{K} c_k (t_k^{(n)} - y_k^{(n)}) . w_{jk} . (h_j)(1 - h_j)(1 - (\frac{e^{a_j} - e^{-a_j}}{e^{a_j} + e^{-a_j}})^2) . x_i$$

because $\frac{\partial h_j}{\partial a_j}$ is nothing but a logistic sigmoid whose parameter is a tanh!

3. Necessary and sufficient condition is that the *gram* matrix $\mathbf{K} = k(\mathbf{x}, \mathbf{x}')$ is positive semidefinite (i.e. $\mathbf{v}^T \mathbf{K} \mathbf{v} \geq 0$ for any choice of $\mathbf{v}$) for all possible choices of the set $\boldsymbol{x}_n$. So if $\mathbf{K}_1$ and $\mathbf{K}_2$ are the *gram* matrices for the valid kernels $k_1(\mathbf{x}, \mathbf{x}')$ and $k_2(\mathbf{x}, \mathbf{x}')$, then, $\mathbf{v}^T (\mathbf{K}_1 + \mathbf{K}_2) \mathbf{v} \geq 0$
4. $k(\mathbf{x}, \mathbf{x}') = exp(k_1(\mathbf{x}, \mathbf{x}')) = \sum_{i=0}^{\infty} \frac{k_1(\mathbf{x}, \mathbf{x}')^i}{i!}$. This is a polynomial in $k_1(\mathbf{x}, \mathbf{x}')$ with positive coefficients. □

3 Graphical Models

Eric Schmidt was so happy when the first problem that Google gave you, got successfully solved by you that he has begun to think that you are *the* smart guy around. Now Eric is being continuously bugged by other corporate CEOs whether Google is using the recent topic models in their search results. Incidentally these people don't know anything about approximate inference and are eager to know why such simple looking graphical models like topic models need approximation in the first place? In his next meeting, Eric has to explain to them very tersely about the difficulty in finding the probability of the corpus given the parameters under the original graphical model. Over

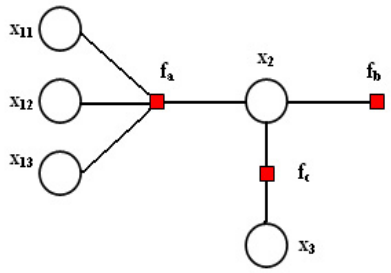

Figure 2: Belief Propagation in Factor graph

a warm and cosy dinner Eric asks you to write down a equation and explain in very few sentences why is it so and give it to him the next day so that he can read in the car while going over to the CEO meeting (yes.. he is a busy man!).

So the first thing that you did was to draw a graphical structure like the one in Fig 1. You next wrote down the following distributions to convert the graphical structure to a model thereby making the model explain how a document can be generated with documents being distributions over latent "topics" and topics being distributions over words in the entire vocabulary.

For each document $d \in 1, ..., M$
Choose a topic proportion $\boldsymbol{\theta}|\boldsymbol{\alpha} \sim Dir(\boldsymbol{\alpha})$
For each of the n words w_n in document d
Choose topic indicator $z_n|\theta \sim Mult(\boldsymbol{\theta})$
Choose a word $w_n|z_n = k, \beta \sim Mult(\beta_{z_n})$
end
end

where you have a vocabulary V of all words in the corpus consisting of M documents. Each document d is a V dimensional vector $\mathbf{w}_m$ and $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ are K-dimensional vectors where K is the number of laten topics. Each topic itself is a multinomial over all words V. Thus β_{ij} is the probability that the observed word w_n [where $n(\in \{1, ..., N_d\})$ is one among $j(\in \{1, ..., V\})$ in the d^{th} $(d \in \{1, ..., M\})$ document] belongs to topic i. We assume there are N_d words in a document d.

Now you need to write down:

1. (4 marks) What are the parameters of the model and why? What are the hidden variables and why?
2. (2 marks) What are the conditional independence statements that hold (or do not hold) for any combination of hidden variables/parameters given the observed word $\mathbf{w}$ in the graphical model obtained by using d-separation (you can also use Bayes Ball)
3. (4 marks) Write a single expression for $p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})$ to show why the likelihood of a document is intractable to compute (and hence exact posterior inference over hidden variables is also intractable)? [HINT: Use factorization from the graph structure and get the marginal/conditional by integration and/or summation of the hidden variables and show by inspection why this integration may be intractable]

For the factor graph in fig. 2

1. (2 marks) Write down the expression for the normalized marginal for node x_2
2. (2 marks) What message is being sent from x_2 to the factor node f_c. If $(x_{11}, x_{12}, x_{13}, x_2)$ are binary variables and F_a is the function defined over factor node f_a, what is the cardinality of the domain of F_a?
3. (2 marks) What is the main difference between sum-product and max-product algorithms?
4. (2 marks) What is the max-sum algorithm? Name an algorithm for belief propagation in factor graphs having loops.
5. (2 marks) Write down the expression for the maximal joint probability of $(x_{11}, x_{12}, x_{13}, x_2, x_3)$

Solution

1. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ because they don't grow with the data

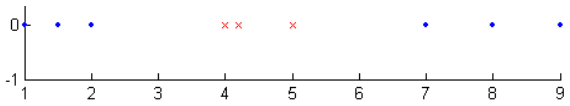

Figure 3: Binary classification

2. $\boldsymbol{\theta}$ and $\mathbf{z}$ because they grow with documents and words respectively
3. $$p(\mathbf{w}|\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{\Gamma(\sum_{j=1}^{V}\alpha_j)}{\prod_{j=1}^{V}\Gamma(\alpha_j)}\int d\boldsymbol{\theta}\,(\prod_{i=1}^{K}\theta_i^{\alpha_i-1})(\prod_{n=1}^{N}\sum_{i=1}^{K}\prod_{j=1}^{V}(\theta_i.\beta_{ij})^{w_n^j})$$

   $w_n^j = 1$ if $n == j$ and 0 otherwise. Since θ and β are coupled, the above integral is intractable.

[Solution for Belief Propagation is straight from the book]

4 Small Questions

1. Harish has asked you to classify the o's (belonging to class 1) and x's (belonging to class 2) data points as given in Fig 3 by a linear decision boundary. What would you do?
   - say that you can do it right in the input data space and show him the plotted decision boundary. Plot a boundary in this case.
   - If Harish is not fully satisfied with the quality of your decision boundary, you need to apply a K-trick (short for "Klever"-trick) to show how you can apply a function on your data and still get an optimal linear decision boundary. Guess a suitable function and plot your favorite transformation and the optimal decision boundary.

   [Solution] Not possible to do it in input space with 100% training accuracy. Just use square of the data ($f(x) = x^2$) and get a linear boundary.