Department of Mathematics and Computer Science
University of Southern Denmark, Odense

March 1, 2013
Marco Chiarandini

DM825 - Introduction to Machine Learning

Sheet 11, Spring 2013

Exercise 1 – Probability theory

Prove the following rule:

$$p(x_i|x_{-i}) = \frac{p(x_1, \ldots, x_N)}{\int p(x_1, \ldots, x_N)dx_i}$$

where $x_{-i} = \{x_1, \ldots, x_N\} \setminus x_i$.

Solution By the product rule

$$p(x_1, \ldots, x_N) = p(x_i|x_{-i})p(x_{-i})$$

Rearranging and marginalizing:

$$\begin{aligned} p(x_i|x_{-i}) &= \frac{p(x_1, \ldots, x_N)}{p(x_{-i})} \\ &= \frac{p(x_1, \ldots, x_N)}{\int p(x_1, \ldots, x_N)dx_i} \end{aligned}$$

Exercise 2 – Naive Bayes

Consider the binary classification problem of spam email in which a binary label $Y \in \{0, 1\}$ is to be predicted from a feature vector $X = (X_1, X_2, \ldots, X_n)$, where $X_i = 1$ if the word i is present in the email and 0 otherwise. Consider a naive Bayes model, in which the components X_i are assumed mutually conditionally independent given the class label Y.

a Draw a directed graphical model corresponding to the naive Bayes model.

Solution

b Find a mathematical expression for the posterior class probability $p(Y = 1|x)$, in terms of the prior class probability $p(Y = 1)$ and the class-conditional densities $p(x_i|y)$.

Solution

$$\begin{aligned} p(Y = 1|x) &= \frac{p(\vec{x}|Y = 1)p(Y = 1)}{p(\vec{x})} \\ &= \frac{\Pi_{i=1}^{n} p(x_i|Y = 1)p(Y = 1)}{\sum_{y=0,1} \Pi_{i=1}^{n} p(x_i|Y = y)p(Y = y)} \end{aligned}$$

c Make now explicit the hyperparameters of the Bernoulli distributions for Y and X_i. Call them, μ and θ_i, respectively. Assume a beta distribution for the prior of these hyperparameters and show how to learn the hyperparameters from a set of training data $(y^j, \vec{x}^j)_{j=1}^m$ using a Bayesian approach. Compare this solution with the one developed in class via maximum likelihood.

Solution

Solution

The hierarchical model is represented in the figure.

For Y we assume

$$p(Y=1|\mu) = \text{Bern}(\mu) = \mu$$

For X_i we the distribution depends by the parent and we assume

$$p(X_i = 1|Y=1, \theta_{i1}) = \text{Bern}(\theta_{i1}) = \theta_{i1}$$
$$p(X_i = 1|Y=0, \theta_{i0}) = \text{Bern}(\theta_{i0}) = \theta_{i0}$$

The prior distribution on the θs and μ captures the uncertainty on these parameters. Assuming a *beta distribution* and referring by θ to both the θ_{iy}s and μ

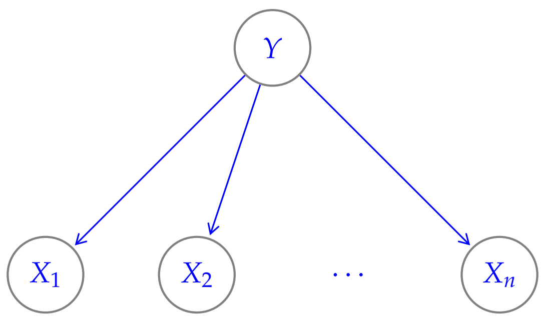
$$p(\theta) = \text{Beta}(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

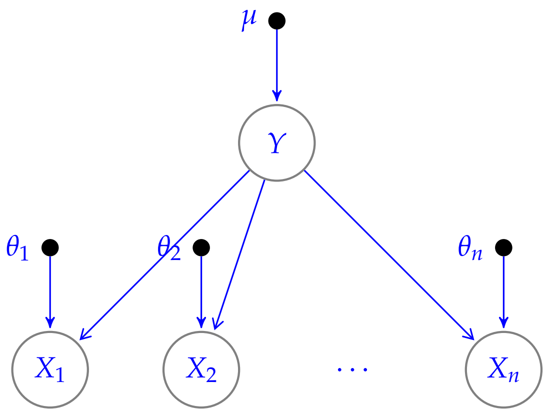
The Gamma function $\Gamma(\cdot)$ is a normalizing function. The parameters α and β with $\alpha > 0$ and $\beta > 0$ are *hyperparameters* of the prior distribution. The mean of a beta distribution is $E[\theta] = \frac{\alpha}{\alpha+\beta}$.

The beta distribution has the *conjugacy property*, that is, the posterior distribution has the same functional form as the prior. This property is convenient because the posterior can be derived in closed form. For the Y node:

$$\begin{aligned} p(\mu|\mathbf{d}) &= \frac{p(\mathbf{d}|\mu)p(\mu)}{p(\mathbf{d})} \\ &= \frac{\text{Bin}(s|\mu)p(\mu)}{p(\mathbf{d})} \\ &\propto \text{Beta}(\mu|\alpha+s, \beta+(m-s)). \end{aligned}$$

where s are the cases of m with $Y=1$.





For the X_i nodes assuming the independence

$$p(\vec{\theta}|\mathbf{d}) = \Pi_{j=1}^{m}\Pi_{y=0,1}p(\theta_{jy}|\mathbf{d})$$

and

$$p(\theta_{i1}|d) = \text{Beta}(\mu|\alpha + s_{i1}, \beta + (s - s_{i1}))$$

where s_{i1} is the number of cases in $\mathbf{d}$ with $X_i = 1$ and $Y = 1$ and s is the number of cases in $\mathbf{d}$ with $Y = 1$.

Thus the prediction for each variable after learning occurred is given by

$$p(Y = 1|\mathbf{d}) = \sum p(Y = 1|\mathbf{d})p(\mu|\mathbf{d}) = E_{p(\mu|\mathbf{d})}[\mu|\mathbf{d}] = \frac{\alpha + s}{\alpha + \beta + m}$$

$$p(X_i = 1|Y = 1, \mathbf{d}) = \sum p(X_i = 1|Y = 1, \mathbf{d}, \theta_{i1})p(\theta_{i1}|\mathbf{d}) = E_{p(\theta_{i1}|\mathbf{d})}[\theta_{i1}|\mathbf{d}] = \frac{\alpha + s_{i1}}{\alpha + \beta + s}$$

This is very similar to what we saw in class derived from the joint likelihood:

$$\begin{aligned}
\phi_y &= \frac{\sum_{j=1}^{m} I\{Y^j = 1\}}{m} = \frac{s}{m} \\
\phi_{i|Y=1} &= \frac{\sum_{j=1}^{m} I\{X_i^j = 1, Y^j = 1\}}{\sum_{j=1}^{m} I\{Y^j = 1\}} = \frac{s_{i1}}{s}
\end{aligned}$$

If we want to predict Y given $\vec{x}$ then we use:

$$\begin{aligned}
p(Y = 1|x, \mathbf{d}) &= \frac{p(\vec{x}|Y = 1, \mathbf{d})p(Y = 1, \mathbf{d})}{p(\vec{x}, \mathbf{d})} \\
&= \frac{\Pi_{i=1}^{n} p(x_i|Y = 1, \mathbf{d})p(Y = 1, \mathbf{d})}{\sum_{y=0,1} \Pi_{i=1}^{n} p(x_i|Y = y, \mathbf{d})p(Y = y, \mathbf{d})}
\end{aligned}$$

Exercise 3 – Directed Graphical Models
Consider the graph in Figure left.

- Write down the standard factorization for the given graph.

  Solution The standard factorization for any directed graphical model can be written as $p(x) = \Pi_{v \in V} p(x_v | x_{pa(v)})$, where $x_{pa(v)}$ are the nodes parent of x_v. Here, this yields

  $p(x) = p(x_1)p(x_2)p(x_3|x_{10})p(x_4|x_2, x_6, x_7)p(x_5|x_9)p(x_6|x_1, x_2)p(x_7)p(x_8)p(x_9|x_3, x_7, x_8)p(x_{10}|x_3).$

- For what pairs (i, j) does the statement X_i is independent of X_j hold? (Don't assume any conditioning in this part.)

  Solution

  The goal is to find all pairs (i, j) such that Xi and Xj are independent. We can achieve this by computing from each node its reachability, that is, the nodes that are reachable by a path that does not have head-to-head subcomponents. From node 1 we can get to nodes 6 and 4. From node 2 we can reach nodes 6, 4, 10, 3, 9, and 5. From nodes 3 and 10 we can reach the sames nodes as node 2. From node 4 we can reach every node but node 8. From node 5 we can reach every node but node 1. From node 6 we can reach any node but nodes 7 and 8. From node 7 we can reach node 9, 4, and 5. Node 8 can only reach nodes 9 and 5. Node 9 can't reach node 1. Finally, node 10 can't reach nodes 1, 7, and 8. Thus (1, 2), (1, 3), (1, 5), (1, 7), (1, 8), (1, 9), (1, 10), (2, 7), (2, 8), (3, 7), (3, 8), (4, 8), (6, 7), (6, 8), (7, 8), (7, 10), and (8, 10) are all independent pairs. In all there are 17 distinct pairs.

- Suppose that we condition on $\{X_2, X_9\}$, shown shaded in the graph. What is the largest set A for which the statement X_1 is conditionally independent of X_A given $\{X_2, X_9\}$ holds? **Solution** We say that $\mathbf{X} \perp \mathbf{Y} \,|\, \mathbf{Z}$ if $\mathbf{X}$ and $\mathbf{Y}$ are d-separated given $\mathbf{Z}$

  in the digraph, that is, if there is no *active* path between any node $X \in \mathbf{X}$ to $Y \in \mathbf{Y}$ given $Z \in \mathbf{Z}$. In class we defined the four conditions for a path to be *active*.

  Checking d-separation implies checking all paths from a vertex to another. This maybe exponential. The following is a linear time algorithm for d-separation. We begin by traversing the graph bottom up, from the leaves to the roots, marking all nodes that are in Z or that have descendants in Z. Intuitively, these nodes will serve to identify a head-to-head structure, ie., $X \to Z \leftarrow Y$. In the second phase, we traverse breadth-first from X to Y, stopping the traversal along a path when we get to a blocked node. A node is blocked if: (a) it is in the "middle" node of a structure $X \to Z \leftarrow Y$ and unmarked in phase I, or (b) is not such a node and is in Z. If our breadth-first search gets from X to Y, then there is a path between them through Z.

  Conditioned on $\{X_2, X_9\}$ there is no active path to nodes 3, 10, 7, 8 and 5. Hence, $A = \{3, 5, 7, 8, 10\}$. Note that nodes 2 and 9 are not elements of the set A because we are conditioning on them.

- What is the largest set B for which X_8 is conditionally independent of X_B given $\{X_2, X_9\}$ holds? **Solution** Conditioned on $\{X_2, X_9\}$ starting at node 8 we cannot

  reach with an active path nodes 1, 5, and 6. Therefore, $B = \{1, 5, 6\}$.

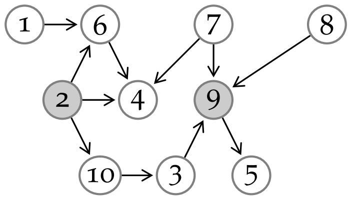


Figure 1: A directed graph.

- Suppose that I wanted to draw a sample from the marginal distribution $p(x_5) = \Pr[X_5 = x_5]$. (Don't assume that X_2 and X_9 are observed.) Describe an efficient algorithm to do so without actually computing the marginal. **Solution**

  We wish to generate a sample of x_5 from the marginal distribution $p(x_5)$. We can achieve this by sampling from the join distribution $p(\vec{x})$ and then marginalizing. For example, given a joint distribution $\Pr[x1, x2]$, one can generate a sample from the marginal distribution of x_1 by sampling from the joint distribution and discarding x_2. To see this, let A be the event that the sample $\bar{x}_1$ lies in some set F. Therefore, $\Pr[A] = \Pr[x_1 \in F \cup x_2 \in \mathbb{R}] = \Pr[x_1 \in F]$. Thus, $\bar{x}_1$ and x_1 have the same distribution.

  Hence we can avoid unnecessary computations applying the following algorithm:

  - calculate the Topological order of the graph
  - sample using the factorization and the topological sorting until you sample x_5.

  Hence, we can first generate a sample of x_2, x_7, and x_8. Then, using factorization, we can generate a sample of from the distribution $p(x_{10}|x_2)$, followed by a sample from the distribution $p(x_3|x_{10})$ by using the sample obtained of x_{10}. Next, we can generate a sample of x_9 from the distribution $p(x_9|x_7, x_8, x_2)$. Finally, we can obtain a sample for x_5 by sampling from the distribution $p(x_5|x_9)$. We can immediately see that generating a sample of x_5 did not require actually sampling from x_1, x_6, or x_4 because conditioned on nodes 2, 7, and 8, x_5 is independent of nodes 1, 6, and 4. Thus, what we've done is generating a sample from the joint distribution $p(x_2, x_3, x_{10}, x_7, x_8, x_9, x_5) = p(x_2)p(x_7)p(x_8)p(x_{10}|x_2)p(x_3|x_{10})p(x_9|x_3, x_8, x_7)p(x_5|x_9)$.