Statistical Methods: Portfolio 4

By Henry Bourne

1 Capturing Dependency of Data Using Graphical Models

Firstly we will ask the question: How does the dependencies between random variables affect modelling the likelihood?

If we assume our data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid then the likelihood factorizes: $p(\boldsymbol{x}_1, ..., \boldsymbol{x}_n|\theta) = \prod_{i=1}^n p(\boldsymbol{x}_i|\theta)$. However, what if we have complicated dependencies in our data? how would we factorize our likelihood then? To solve this problem we can first convert our dependencies into a graphical representation and then use the graph to guide our factorization; this is called **graphical modelling**.

Let X and Y be two random variables, if $p(X, Y) = p(X)p(Y)$ [1] then $X \perp Y$ which denotes that X and Y are **independent**. We say X is **conditionally independent** of Y given Z, denoted $X \perp Y|Z$, if $p(X, Y|Z) = p(X|Z)p(Y|Z)$ or equivalently $p(X, Y, Z) \propto P(X, Z)P(Y, Z)$ (which note is a factorization) [2].

Conditional independence and independence tell us how information is exchanged between RVs, ie. $X \perp Y$ tells us no information exchanges between X and Y and $X \perp Y|Z$ tells us no direct information exchanges between X and Z.

Given many RVs, listing all their dependencies can be cumbersome, let's instead use a graphical representation. We let each RV be a node in the graph and join two nodes if they are not dependent on each other, ie. if $X \perp Y$ then we do not draw an edge between X and Y and if $X \perp Y|Z$ then we only draw edges between "Z and X" and "Z and Y". We can read from the graph (or conversely construct it) by checking which RV another RV isn't directly connected to (or equivalently is (conditionally) independent to) and then following a path that connects the two (or equivalently check what RVs are being conditioned on) we can read the following conditional independency: <Node> $\perp$ <Node/s to which it doesn't have an edge/s> | <Nodes on the path> (or equivalently we draw a edge from the node to the start of the path along the path and from the node at the end of the path to the nodes that represent the RV/s being conditioned on).

Can we represent a probability distribution factorization using a graph? Given a graph $G = (E, V)$, we say $p(X)$ factorizes over G if $p(X) \propto \prod_{c \in C} g_c(X^c)$. Where C is the set of all cliques in G and g_c is a function defined on $X^{(c)}$ which is the subset of X restricted on c. A **clique** is a fully connected subgraph.

It turns out that these two graphical representations are equivalent! ie. if p factorizes over G, p satisfies all conditional independence represented by G and vice-versa. We verify this using an example in question A.1.1.

2 Markov Network (Undirected Graph)

We will first start with a definition:

Definition 2.1 ***Markov Network / Undirected Graphical Model:***
An undirected graph representing the conditional independence of the probability distribution $p(X)$

More specifically we can define the **Gaussian Markov Network** (GMN) which uses an undirected graph to represent the conditional independence of a random variable that is multivariate gaussian. Let $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{x} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ then $p(\boldsymbol{x}) \propto \exp(-\frac{1}{2}\boldsymbol{x}(\boldsymbol{\Sigma})^{-1}\boldsymbol{x}^T)$. Let $\boldsymbol{\Theta} = (\boldsymbol{\Sigma})^{-1}$, then we can rewrite this as $p(\boldsymbol{x}) \propto \exp(-\frac{1}{2}\sum_{u,v} \boldsymbol{\Theta}^{(u,v)} x^{(u)} x^{(v)}) \propto \prod_{u,v;\boldsymbol{\Theta}^{(u,v)} \neq 0} \exp(-\boldsymbol{\Theta}^{(u,v)} x^{(u)} x^{(v)})$. Hence, we can write:

$$p(\boldsymbol{x}) \propto \prod_{u,v;\boldsymbol{\Theta}^{(u,v)} \neq 0} g_{u,v}(x^{(u)}, x^{(v)}) \tag{1}$$

[1] Equivalently $p(X|Y) = p(X)$ and $p(Y|X) = p(Y)$ also define independence of X and Y

[2] Further equivalent definitions are that $p(X|Y, Z) = p(X|Z)$ and $p(Y|X, Z) = (Y|Z)$

So, $p(\boldsymbol{x})$ factorizes over G, where we can define G using the following adjacency matrix:

$$A^{(u,v)} = \begin{cases} 0, & \Theta^{(u,v)} = 0 \\ 1, & \Theta^{(u,v)} \neq 0 \end{cases} \tag{2}$$

Note that G must be undirected as $\boldsymbol{\Sigma}$ and therefore Θ are symmetric so if $\Theta^{(u,v)} \neq 0$ then $\Theta^{(v,u)} \neq 0$ and therefore every edge in G is bidirectional. Also note that given a graph G we can decode from it the sparsity of $\boldsymbol{\Theta}$, ie. all the entries that are zero. This means that we can go from the dependencies to a graph and then from the graph find the factorization. Even if we do not know the conditional independence of $p(\boldsymbol{x})$, we can still find a factorization of $p(\boldsymbol{x})$ as follows: given a dataset D we can fit a $\hat{\boldsymbol{\Theta}}$ (for example using the MLE), the sparsity of $\hat{\boldsymbol{\Theta}}$ then gives a graph corresponding to the factorization of $p(\boldsymbol{x})$, additionally this graph also gives us the conditional independencies of the random variables.

2.1 Conditional Markov Network

In many tasks we may want to find the conditional distribution, $p(Y|X)$. So, how do we factorize a conditional distribution over G?

Definition 2.2 ***Conditional Markov Network:***
We say a conditional probability distribution $P(Y|X)$ factorizes over G, whose nodes $V = X \cup Y$, if:

$$p(Y|X) = \frac{1}{N(X)} \prod_{c \in C} g_c(V_c) \tag{3}$$

Where $C := \{c \text{ is a clique in } G | V_c \nsubseteq X\}$ and the normalizing constant $N(X) := \int \prod_{c \in C} g_c(V_c) dY$

Note that $p(Y|X)$ does not include factors defined on subsets of conditioning variable X, eg. if $p(Y|X) = \frac{1}{N(X)} g_1(Y,X) g_2(X)$ then $N(X) = \int g_1(Y,X) g_2(X) dY = g_2(X) \int g_1(Y,X) dY$ and therefore $p(Y|X) = \frac{g_1(Y,X) g_2(X)}{g_2(X) \int g_1(Y,X) dY} = \frac{g_1(Y,X)}{\int g_1(Y,X) dY}$.

2.1.1 Logistic Regression

This way of constructing a conditional likelihood gives us a logistic regression. Let's consider a simple Markov Network where we have $Y \in \{-1, 1\}$ and $X \in \mathbb{R}^d$ where we draw an edge between each node $X^{(i)}$ and Y. The factorization we obtain from this graph is $p(Y|X) = \frac{1}{N(X)} \prod_i g_i(Y, X^{(i)})$ with $N(X) = \sum_{Y \in \{-1,1\}} \prod_i g_i(Y, X^{(i)})$.

Let us now construct a model of $p(Y|X)$ by setting: $g_i(Y = y, X^{(i)} = x^{(i)}; \beta_i, \beta_0) := \exp(y(\beta^{(i)} \cdot x^{(i)} + \beta_0))$. Then our model is $p(y|\boldsymbol{x}; \boldsymbol{\beta}, \beta_0) = \frac{1}{N(X)} \prod_i \exp(y(\beta^{(i)} \cdot x^{(i)} + \beta_0)) = \frac{1}{N(X)} \exp(y(\langle \boldsymbol{\beta}, \boldsymbol{x} \rangle + d\beta_0))$, where $N(X; \boldsymbol{\beta}, \beta_0) = \sum_{y \in \{1,-1\}} \exp(y(\langle \boldsymbol{\beta}, \boldsymbol{x} \rangle + d\beta_0))$. We can then fit $\boldsymbol{\beta}, \beta_0$ using the MLE (question A.2.2 shows this is the same logistic regression we discussed in the previous portfolio).

3 Bayesian Network (Directed Graph)

So far we have been working with undirected graphs to represent conditional independencies. However, what if we want to represent causal relationships for example? then what might serve as a better model would be a directed graphical model, we will be using a:

Definition 3.1 ***Directed Acyclic Graph (DAG):***
A graph, $G = (V, E)$, where E is a directed edge set and G is an acyclic graph, ie. it has no directed cycles.

We can represent the factorization of a probability distribution using a DAG. We say a probability distribution $p(X)$ factorizes over a DAG G if $p(X) = \prod_{v \in V} p(X_v | X_{\text{parent}(X_v)})$. And we can construct the DAG using a similar methodology to constructing undirected graphical models.

We can also represent conditional independencies using a DAG. Given a DAG G, X_v is independent of $X_{\text{non-desc}(X_v)}$ given $X_{\text{parent}(X_v)}, \forall v$. This is analogous to the markov net as X_v and all non-descendants of X_v are "made independent" by the parents of X_v. Also note that knowing

$X_{\text{parent}(X_v)}, X_{\text{non-desc}(X_v)}$ tell us nothing new about X_v. Just as in section 1 we can show that the graph given by the dependencies and the graph given by the factorization are equivalent.

We can then define a:

Definition 3.2 ***Bayesian Network:***
A DAG G constructed using the factorization of a probability distribution, $p(\boldsymbol{x})$.

Lets consider again the simple graph described in section 2.1.1 except we have a directed edge (instead of undirected) going from Y to each $X^{(i)}$. We can write down the conditional probability from this graph:

$$P(Y|X) = \frac{\prod_i P(X^{(i)}|Y)P(Y)}{P(X)} \tag{4}$$

Note that this is how Naive Bayes is derived.

3.1 Bayesian Network for Classification vs. Logistic Regression

Consider the setup with the simple graph we just introduced, we note that it has the same structure as the graph that lead to a logistic regression, except the edges were directed. What are the similarities and differences between naive Bayes and logistic regression?

First lets consider the factorization, for both we have pairwise factors between Y and the $X^{(i)}$. However, in Naive Bayes the factors are conditional probabilities as opposed to cliques in logistic regression. Next, we have that for the probabilistic model they both use $p(Y|X)$ to make a prediction, however, Naive Bayes does not give you $p(Y|X)$ it only give you it up to a constant. Next, for the training/fitting of the classifier in the logistic regression case we estimate $p(Y|X)$ as opposed to $P(X|Y)$ in the case of naive bayes. Finally, for both we use the prediction rule: $\hat{y} := \text{argmax}_y\, p(Y|X)$.

Appendices

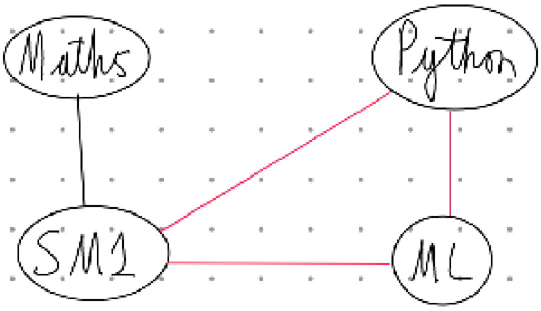
A Homeworks

A.1 For Section 1

Question A.1.1 ***Show equivalence between the factorization and conditional independence over G in the Scores of units example:***
First we will create G using the following factorization:

$$p(\mathit{Maths}, \mathit{SM}1, \mathit{Python}, \mathit{ML}) \propto g_1(\mathit{Maths}, \mathit{SM}1) \cdot g_2(\mathit{Python}, \mathit{ML}, \mathit{SM}1) \tag{5}$$

The graph corresponding to this factorization is: Where in black is the edges corresponding to the clique given by the first factor and in red the edges corresponding to the clique given by the second factor. The conditional independencies encoded by this graph are:



1. *Maths* $\perp$ *Python* | *SM1*
2. *Maths* $\perp$ *Python* | *SM1, ML*
3. *Maths* $\perp$ *ML* | *SM1*
4. *Maths* $\perp$ *ML* | *SM1 , Python*
5. *Maths* $\perp$ *ML, Python* | *SM1*

Which are all the conditional independencies of $p(\mathit{Maths}, \mathit{SM}1, \mathit{Python}, \mathit{ML})$.
Now we will create G using all the conditional independencies, which are:

1. *Maths* $\perp$ *Python* | *SM1*
2. *Maths* $\perp$ *Python* | *SM1, ML*
3. *Maths* $\perp$ *ML* | *SM1*
4. *Maths* $\perp$ *ML* | *SM1 , Python*
5. *Maths* $\perp$ *ML, Python* | *SM1*

The graph we get is the same as before, and reading the graph its clear that it encodes the factorization of $p(\mathit{Maths}, \mathit{SM}1, \mathit{Python}, \mathit{ML})$. *Hence, shown.*

A.2 For Section 2

Question A.2.1 ***Suppose graph G encodes all conditional independencies in your Gaussian distribution*** $p(x)$***. Let's say G contains 3 edges and 5 nodes. How many non-zero elements are there in inverse covariance matrix of p?:***
There are 25 entries in $\boldsymbol{\Theta}$ *and in the adjacency matrix of G (also with 25 entries) there are 6 non-zero (ie. equal to 1) entries hence only 6 entries in* $\boldsymbol{\Theta}$ *that are non-zero (from eq. (2)).*

Question A.2.2 ***In section 2.1.1 we constructed a logistic regression from our simple Markov network model where $\hat{\boldsymbol{\beta}}, \hat{\beta}_0 = \mathrm{argmax}_{\boldsymbol{\beta},\beta_0} \sum_{i=1}^n \log(p(y_i|\boldsymbol{x}_i;\boldsymbol{\beta},\beta_0))$ show that this is the same logistic regression we talked about in portfolio 3:***
Note that we can write:

$$p(y=-1|\boldsymbol{x}) = 1/(1+\frac{p(\boldsymbol{x}|y=+1)p(y=+1)}{p(\boldsymbol{x}|y=-1)p(y=-1)}) \tag{6}$$

And for $p(y=-1|\boldsymbol{x})$ the same is true but with the inverse of the ratio of the densities. We can rewrite this more generally as:

$$p(y|\boldsymbol{x};\boldsymbol{\beta},\beta_0) = \sigma(f(\boldsymbol{x};\boldsymbol{\beta},\beta_0)\cdot y) \tag{7}$$

where $f(\boldsymbol{x};\boldsymbol{\beta},\beta_0) = log([p(\boldsymbol{x}|y=+1)p(y=+1)]/[p(\boldsymbol{x}|y=-1)p(y=-1)])$. Hence we can rewrite our MLE as the logistic regression:

$$\hat{\boldsymbol{\beta}}, \hat{\beta}_0 = \underset{\boldsymbol{\beta},\beta_0}{\mathrm{argmax}} \sum_{i=1}^n \log(\sigma(f(\boldsymbol{x}_i;\boldsymbol{\beta},\beta_0)\cdot y_i)) \tag{8}$$

Which is the same logistic regression we had in portfolio 3.

A.3 For Section 3

Question A.3.1 ***Given the simple Bayesian Network model described in section 3, however, now with one additional node X' which has one inbound directed edge from $X^{(1)}$. Given this Bayesian Network for a classification task, should you include feature X' for classification? and why?***
To be able to solve the classification problem we would like to find $p(Y|X)$ which for this Bayesian Network is equal to:

$$P(Y|X) = \frac{\prod_i P(X^{(i)}|Y)P(Y)P(X'|X^{(1)})}{P(X)} \tag{9}$$

Hence, we should include feature X' for classification as our factorization and therefore prediction depends on it.