

Note Set 2, Multivariate Probability Models:

© Padhraic Smyth, 2009

1 Introduction

This set of notes will cover basic concepts in **multivariate probability models**, i.e., probability models involving multiple random variables. We will begin by discussing the important concepts of independence and conditional independence and then look at the general framework of graphical models. We will then look at specific examples of models: for joint distributions we will look at the naive Bayes probability model and Markov models, and for joint densities we will look at the multivariate Gaussian model. Note that the ideas we discuss below are applicable equally well to both discrete-valued and continuous variables (and their combinations), but we will primarily focus on discrete-valued variables in our examples.

2 Independence and Conditional Independence

2.1 Independence

Let A and B be two random variables (we can assume they are discrete to be concrete, but they could also be continuous).

Definition 1 A and B are **independent** iff $P(a, b) = P(a)P(b)$, $\forall a, b$.

We see that independence of two random variables implies that their probability distribution **factors**. For example, if A and B each took 100 values, then the full joint distribution $P(a, b)$ would consist of a table of $100 \times 100 = 10,000$ probabilities. However, if one assumes or knows that A and B are independent, then we can factor the table as a product of $P(a)$ and $P(b)$ and only need $100 + 100 = 200$ probabilities which is far fewer! Of course what we save in terms of complexity

(of the model) is offset by the fact that we won't be able to model all of the possible dependencies between a and b values as we might if used a full unconstrained joint distribution $P(a, b)$ without any independence assumption.

It is important to understand that independence is often **assumed** for convenience in modeling problems. We are all familiar with simple games of chance, such as tossing a coin or throwing a die, where each toss or throw can be assumed to be independent of all previous tosses or throws. However, in other problems, many phenomena are dependent in some subtle way or another, e.g., even the temperature in Irvine tomorrow may be related to the likelihood of rainfall in Beijing today. In practice, however, we can build useful models that assume that variables are independent even when they are slightly dependent in reality. Unlike the properties of probability models that were discussed in Note Set 1, independence is **an assumption** that we can make in building a probability model, not a basic property.

We can extend the definition of independence to more than 2 random variables. For example, we say that A, B, C are mutually independent if $P(a, b, c) = p(a)p(b)p(c) \forall a, b, c$. Is independence transitive? i.e., if A is independent of B , and B is independent of C , then is A independent of C ? (Homework Problem).

There is another way to define independence using conditional probability that provides some more intuition as to what independence means. The definition of independence we provided above is mathematically equivalent to the following alternate definition:

Definition 2 A and B are independent iff $P(a|b) = P(a)$ and $P(b|a) = P(b) \forall a, b$.

The condition $P(a|b) = P(a)$ can be interpreted as stating that the event $P(B = b)$ provides no information about the event $P(A = a)$, i.e., our belief in a is not changed at all if we learn that $B = b$.

2.2 Conditional Independence

A more general notion of independence is **conditional independence**, defined as follows:

Definition 3 A and B are conditionally independent given random variable C iff $P(a, b|c) = P(a|c)P(b|c), \forall a, b, c$.

Again this can be equivalently stated as

$$P(a|b, c) = P(a|c) \quad \text{or} \quad P(b|a, c) = P(b|c).$$

We can interpret this first equation as saying that if we already know $C = c$, then learning that $B = b$ does not change our belief in $A = a$, i.e., knowing b does not change our belief in a if we already know c .

This type of **conditional** independence is very useful in practical modeling problems. For example, we might assume that two disease symptoms A and B are conditionally independent given the disease C , e.g., the occurrence of headache and fever are conditionally independent given flu. As we mentioned earlier, these independence relations are not necessarily true in the real-world, but we can often **assume** that they hold for the purposes of building an **approximate model**—this is somewhat analogous to approximating a non-linear function with a “linear first-order approximation” in applied mathematical modeling.

EXAMPLE 1: Rainfall in Two Parts of the World: Let X_t be the maximum daily temperature in Irvine and let Y_t be the maximum daily temperature in Beijing, both measured on day t . Intuitively we don’t believe that X_t and Y_t should directly depend on each other. However, if we have no other variables in our model and we wanted to build an accurate model for the joint density $p(x_t, y_t)$ we would in fact need to assume that they directly depend on each other. For example, they will both tend to rise and fall in a similar manner depending on the time of year, given that they are both in the Northern hemisphere—so learning that it is relatively cool in Irvine will cause you to update your degree of belief about the temperature in Beijing that day.

What’s really going on here is that Irvine and Beijing are “connected” indirectly by large-scale processes governing seasonal climate. For example, if we had a 3rd variable Z that represents the day of the year in the Western calendar (taking values 1 through 365), we could use the conditional independence model $p(x, y, z) = p(x|z)p(y|z)p(z)$ and assume that Irvine and Beijing are conditionally independent if we know what day of the year it is. This is a much more sensible model than the one that has to connect Irvine and Beijing directly together. We can start to see here how we will be able to simplify models by introducing “key variables” that induce conditional independence relations among large sets of other variables (e.g., imagine building a model for dozens of cities rather than two). Of course, although we like our conditional independence model better than the other model, it might not capture all aspects of variation between X and Y . If, for example, some years are different to others in the Northern hemisphere temperature patterns (for example,

a large volcano eruption like Mount Pinatubo can cause systematic changes in temperature over the whole globe) then our conditional independence model would only be an approximation since the actual temperature in Irvine (by being systematically warmer or cooler that year) would carry some information about the probability distribution over temperatures in Beijing, beyond what is captured in the variable Z_t representing the day of the year.

There is a famous quote by the statistician George Box: “*all models are wrong, but some are useful.*” What he is saying here is that any model will not match reality exactly, so in a sense all models are wrong at some level—but nonetheless, some models are nonetheless very useful even though they don’t capture every detail of the real-world. For example, in physics, Newtonian models of motion don’t take into account relativity, but are still very useful. Similarly we will see with probabilistic modeling that models that use conditional independence assumptions may be over-simplifying certain aspects of the real-world, but may nonetheless be very useful in modeling.

Note that in the model above for temperature, that X and Y are not marginally independent, i.e., $p(x, y) \neq p(x)p(y)$ even though they are conditionally independent. In general, conditional independence does not imply marginal independence. In the example above we can see why this is the case: conditioned on Z (the day of the year), the temperatures in Irvine and Beijing can be assumed to have no information about each other. But if we don’t know Z , then they have information about each other, i.e., they are dependent if Z is unknown.

Note that we can extend the definition of conditional independence to multiple random variables, e.g., we say that A, B, C, D are conditionally independent given Z , if and only if

$$P(a, b, c, d|z) = P(a|z)P(b|z)P(c|z)P(d|z).$$

Thus, if we know the value of Z , then none of A, B, C, D carry any information about each other.

EXAMPLE 2: First-Order Markov Models: Consider a sequence of random variables $X_1, \dots, X_t, \dots, X_T$ where the random variable at position t in the sequence is indexed as X_t . All random variables $X_t, 1 \leq t \leq T$, are assumed to be taking values from the same set (whether discrete or real-valued). As an example consider X_t representing the t th word in a spoken sentence or a text document—in this case the number of possible values for each X_t is the number of words that the model knows about and could be as large as 50,000 or 100,000 (e.g., in real-world speech recognition or text modeling applications).

The sequence $X_1, \dots, X_t, \dots, X_T$ is defined to have first-order Markov property if the following property holds true:

$$P(x_{t+1}|x_t, x_{t-1}, \dots, x_1) = P(x_{t+1}|x_t), \quad 1 \leq t \leq T - 1$$

which is just another way of stating that the variable X_{t+1} is conditionally independent of all variables X_{t-1} back to X_1 , conditioned on X_t . This assumption is sometimes stated as “the future only depends on the present and not on the past.” We can see that for modeling sequences of words that making an assumption such as this one is almost essential from a practical viewpoint since it means we only have to model dependencies between pairs of successive words, and not the joint distribution of long sequences of words (of which there are an exponential number of such sequences as the length grows). However, by only modeling adjacent words we cannot for example model some of the natural structure of language, such as various aspects of grammar and phrasing, which tend to lead to longer-range dependencies. Nonetheless, the simple first-order Markov model is often very useful in practice.

EXAMPLE 3: Higher-Order Markov Models: We can generalize the Markov model above to a k th order Markov model, where we assume that

$$P(x_{t+1}|x_t, x_{t-1}, \dots, x_1) = P(x_{t+1}|x_t, \dots, x_{t-k+1})$$

for some fixed integer $k > 0$. For $k = 1$ we get the first order Markov model. For $k = 2$ we get a 2nd-order Markov model where the next state now depends on the previous *two* states, and so on with 3rd order, 4th order, etc, Markov models. The conditional independence assumption being made is that the next state is conditionally independent of earlier states given the k preceding state values. (Here we are using the word “state” to refer to the value of the variables X_t , a common convention when describing Markov chains). Modeling of DNA sequences in bioinformatics is (for example) one area where such high-order models can be used: in that case, each position X_t can take one of 4 possible values **A**, **G**, **T**, **C**, and we can “afford” to build models that are richer than simple first-order dependency. However, the number of probabilities that we require for our model increases exponentially with k since we need to specify an exponentially increasing number of probabilities as k increases, specifically order of m^{k+1} . Various interesting extensions are possible for problems where m is very large (such as modeling words in text), such as clustering words together into groups and modeling the conditional probabilities at the group level, and also searching for specific combinations of values that have higher order dependency (rather than assuming all combinations have high-order dependency) and then representing these dependencies in a tree-structured Markov model.

2.3 The Naive Bayes Model

The term **naive Bayes model** is sometimes used to refer to a model where we have a set of d random variables¹ X_1, \dots, X_d that are assumed to be conditionally independent given a discrete random variable C . The term “naive Bayes” was originally used for this model to convey the idea that the model is “naive” in terms of its modeling assumptions.

The naive Bayes model is just a type of conditional independence model where we have a number of variables (here, X_1, \dots, X_d) that are assumed to be conditionally independent given another variable C . A typical application is where X_1, \dots, X_d consists of a set of d discrete random variables that we can measure (sometimes called the **features** or **attributes** of an object), with another discrete-valued variable C taking values $\{1, \dots, m\}$ (sometimes called the class variable) that represents a property of the object that is not directly observed and must be inferred. The probabilities of different values of C can be calculated via Bayes rule as follows:

$$\begin{aligned}
 P(C = j|x_1, \dots, x_d) &= \frac{P(x_1, \dots, x_d|C = j)P(C = j)}{P(x_1, \dots, x_d)} \\
 &\quad \text{(by Bayes' rule)} \\
 &= \frac{\left(\prod_{i=1}^d P(x_i|C = j)\right)P(C = j)}{P(x_1, \dots, x_d)} \\
 &\quad \text{(invoking the conditional independence assumption)} \\
 &= \frac{1}{K} \left(\prod_{i=1}^d P(x_i|C = j)\right)P(C = j), \quad 1 \leq j \leq m \\
 &\quad \text{(where } K \text{ is a proportionality constant independent of } j\text{).}
 \end{aligned}$$

Notice that we can compute $P(C = j|x_1, \dots, x_d)$ as being proportional to a product of simpler individual terms $P(x_i|C = j)$ (by virtue of the conditional independence assumption) and $P(C = j)$. Be aware of the “mixing” of notation here: $C = j$ refers to the j th value of class variable C , whereas x_i refers to some value of variable X_i —this overloading of notation is hard to avoid without using superscripts (which brings its own problems in terms of interpretation), but the interpretation should be reasonably clear from the context.

A full joint distribution over X_1, \dots, X_d and C would require order of mK^d probabilities to specify, if each of the X_i variables took K values. For say $d = 100$ this is clearly impractical. On

¹Note the change in notation here where the subscript i in X_i indicates variable i , and x_i represents a possible value for variable X_i (rather than the i th value of variable X , as we had in Note Set 1).

the other hand, with the conditional independence assumption we need far fewer probabilities in total (Homework Problem to compute exactly how many).

EXAMPLE 4: Medical Diagnosis: The naive Bayes model has been widely used in the past for building simple models for automated medical diagnosis. The X_i variables represent symptoms and the C variable represents the disease variable. As well as the advantage of requiring far fewer parameters than a full joint distribution, the naive Bayes model is also quite easy to interpret, since if we write the equations above in log form (by taking the log of each side) we get:

$$\log P(C = j|x_1, \dots, x_d) = \sum_{i=1}^d \log P(x_i|C = j) + \log P(C = j) - \log K.$$

This can be interpreted as a simple **additive model**. We can look at the final results and see which of the terms $\log P(x_i|C = j)$ had the most impact on the final answer. For example, if we measure a value x_i for a feature such that $P(x_i|C = j)$ is close to zero (i.e., according to the model it is highly unlikely that we would observe that value of x_i if the class variable is $C = j$), then $\log P(x_i|C = j)$ will be a very large negative number in this case and this will tend to make $\log P(C = j)$ very negative (and $P(C = j)$ very small) relative to the other possible class values (assuming that the value x_i is not as unlikely conditioned on the other class values). Thus, the contribution of individual features can be interpreted in the final outcome because of the additive nature of the model.

Of course the limitation of the naive Bayes model is the fact that it ignores any interactions among the features and how these interactions can affect the probability of $C = j$. For example, “exclusive-OR” types of relations cannot be modeled, such as a person being likely to have the disease if they have either symptom X_1 or symptom X_2 but being *very unlikely* to have the disease if they exhibit both symptoms at the same time. Keep in mind, however, that if the occurrence of both symptoms together is relatively rare, or if modeling of $P(C = j|...)$ is dominated by other variables, then not modeling such interactions may have relatively little impact overall in terms of the performance of the model overall.

EXAMPLE 5: Spam Email Classification: By far the most ubiquitous application of the naive Bayes model in 2005 is in open-source and commercial spam email filters. Most such systems use a naive Bayes model to classify incoming emails into two classes, “spam” or “not spam”, as represented by a binary class variable C . The feature variables X_1, \dots, X_d represent the occurrence of different words or phrases in the text. For example X_i could represent the presence or absence

of the phrase **free offer** in the email, or the number of times this phrase occurs. Selection of which phrases to include in the model is of course somewhat of an art, but there are techniques that can search for good “discriminative” features to include given a set of training data. Having large sets of training data is of course not a problem given that we are all inundated with both spam and non-spam emails on a daily basis. For real-world spam filters there are of course many other features besides the words in the email that can be taken into account (header information, subject line information, HTML structure, sender address, etc) but even with these additional features the underlying models that are used still tend to be naive Bayes models in practice.

3 Directed Graphical Models or Bayesian Networks

Conditional independence is a very useful general framework for structuring a probability model in terms of how our variables are connected in a model. We can take this idea a bit further and use graph theory to provide a formal modeling language for specifying and computing with sets of random variables that have various dependence and independence relations among them.

3.1 Definition of a Directed Graphical Model

Consider a set of d random variables X_1, \dots, X_d , assumed discrete-valued for now, but all of what we will say below also applies to the continuous or mixed cases.

In a directed graphical model² we represent each random variable in our model by a node in a directed graph. A directed edge exists in the graph between X_i and X_j if X_j directly depends on X_i . Cycles are not allowed in the graph.

A directed graphical model uses a directed graph to represent a specific factorization of the joint distribution as follows:

$$P(X_1, \dots, X_d) = \prod_{i=1}^d P(X_i | \text{parents}(X_i))$$

where $\text{parents}(X_i)$ are the parents of the node X_i in the directed graph, i.e., the set of nodes that point to variable X_i . For example, if we have a graph with an edge from A to C and an edge from B to C , and no other edges, then A and B have no parents in this graph, C has two parents A and

²“Bayesian network” is another name used to describe directed graphical models—they are the same thing.

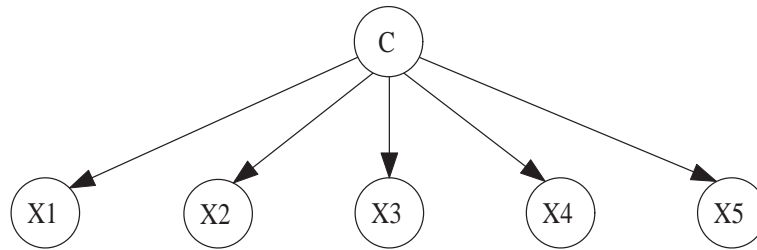


Figure 1: The graphical model representing a naive Bayes model with a class variable C and 5 features X_1, \dots, X_5 .

B , and the joint distribution specified by this graph is

$$\begin{aligned} p(a, b, c) &= P(a|\text{parents}(a)) P(b|\text{parents}(b))P(c|\text{parents}(c)) \\ &= P(a)P(b)P(c|a, b). \end{aligned}$$

Thus, we can “read off from the graph” the joint distribution as a product of variables and their parents.

Note that the graphical model provides a convenient mechanism for describing sets of conditional dependence relations. Loosely speaking, a variable X_i in a directed graphical model is conditionally independent of all non-descendant variables in the graph given the values of its parent variables. There is a direct equivalence between the structure of the graph, the conditional independence relations implied by the graph structure, and the factorization of the joint probability into a product of local probability tables involving a variable and its parents. Given any 1 of these 3 representations of structure we can recover the other 2.

The graph framework is useful because it allows us represent, visualize and communicate the **structure** of a model in a systematic manner. There is also a rich theory that links the computational complexity of performing calculations with the probability model such as marginalization (summing out certain variables) with the intrinsic structure of the graph. For example, graphs that can be represented as trees, or very sparse graphs, tend to be much more efficient for computation versus graphs that have cycles (defined as cycles in the graph that results when we drop directionality of the edges) and/or dense graphs. We will not dwell on the rich theoretical framework of graphical models here but instead use graphical models as a convenient mechanism for describing probability models that have structure in them.

EXAMPLE 6: The Directed Graphical Model for Naive Bayes: The graphical model for



Figure 2: The graphical model for a Markov chain defined on 5 variables.

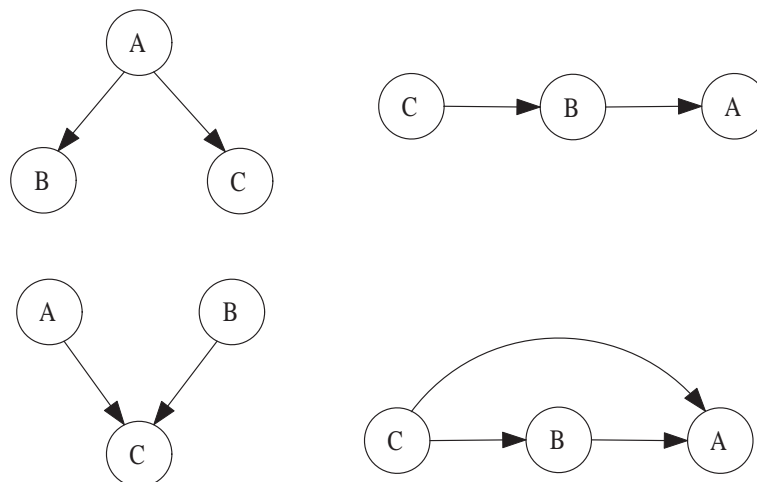


Figure 3: Examples of different graphical models for 3 random variables $A, B,$ and $C.$

naive Bayes is very simple (Figure 1), a single class variable as the single parent of d child nodes, with a directed edge from C to each of $X_1, \dots, X_d.$

EXAMPLE 7: The Directed Graphical Model for a Markov Chain: We defined earlier the Markov chain property for a sequence of random variables $X_1, \dots, X_t, \dots, X_T.$ The graphical model is again very simple (Figure 2), consisting of a graph in the form of a chain with a directed edge from each X_t to X_{t+1} and with X_1 having no parents.

EXAMPLE 8: Directed Graphical Models for 3 Variables: Consider 3 variables $A, B, C.$ We can define a number of different simple graphical models for these three variables as follows (see Figure 3):

- the independence model, $P(a, b, c) = P(a)P(b)P(c)$ consists of the empty graph (since none of the variables have any parents).
- We can have 3 different “naive Bayes”-type conditional independence models, with any 2 of the variables being conditionally independent given the other, and the conditioning variable

being the parent of the other 2.

- We can also have simple Markov chains, e.g., $P(a, b, c) = P(a|b)P(b|c)P(c)$ where C points to B which points to A .
- Another model is the “multiple cause” model, e.g., where the “causes” A and B point to a common “symptom” C . This is useful for example in modeling two medical conditions that can occur independently in a patient, and that have a common symptom.
- The model with no conditional independence relations at all (sometimes referred to as the **saturated model**) can be described in graphical form by recalling the factorization property from Note Set 1 that holds for all distributions and densities, e.g., $P(a, b, c) = P(a|b, c)P(b|c)P(c)$. This is equivalent to a graph where B and C point to A , C points to B , and nothing points to C . We could describe this same saturated model in different ways depending on the ordering. In fact for any arbitrary ordering of the variables the saturated model is equivalent to a graph where each variable has as parents *all* of the variables that precede it in the ordering.

3.2 The Multivariate Gaussian Model

Consider a set of random variables X_1, X_2, \dots, X_d , where x_1, x_2, \dots, x_d will be used to denote generic values for this set of random variables. We will use the shorthand notation $\underline{x} = (x_1, x_2, \dots, x_d)$ to denote a d -dimensional vector of values. The standard convention is to assume that the vector is a column vector, i.e., has dimension $d \times 1$.

The multivariate Gaussian density function is defined as:

$$p(\underline{x}) = p(x_1, x_2, \dots, x_d) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T \Sigma^{-1} (\underline{x}-\underline{\mu})}$$

where $\underline{\mu} = (\mu_1, \dots, \mu_d)$ is a $d \times 1$ dimensional vector of mean values and Σ is a $d \times d$ covariance matrix with entries $\sigma_{ij} = cov(X_i, X_j)$ and diagonal terms $\sigma_{ii} = cov(X_i, X_i)$ being the variance σ_i^2 for each of the individual X_i variables.

Although at first this expression looks rather formidable, it can be broken down into a simpler version, namely

$$p(\underline{x}) = \frac{1}{C} e^{-\frac{1}{2}d_{\Sigma}(\underline{x}, \underline{\mu})}$$

where C is just a normalization constant and

$$d_{\Sigma}(\underline{x}, \underline{\mu}) = (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu})$$

defines a non-negative scalar distance between the vector \underline{x} and the mean $\underline{\mu}$. This distance is a generalization of the standard notion of Euclidean distance. If $\Sigma = I$, the identity matrix, we get $(\underline{x} - \underline{\mu})^T(\underline{x} - \underline{\mu})$, which is the square of the Euclidean distance $\sum_{i=1}^d (x_i - \mu_i)^2$ between \underline{x} and $\underline{\mu}$. If Σ is diagonal, then Σ^{-1} is also diagonal, and the distance becomes a weighted Euclidean distance, $\sum_{i=1}^d w_i (x_i - \mu_i)^2$ where each weight $w_i = \frac{1}{\text{sigma}_i^2}$. Thus, the contribution to the overall distance is scaled by the inverse of the variance in each dimension—this is equivalent to pre-scaling all the X_i variables to have unit variance, so that the contributions to the overall distance from each dimension would have equal weight from a variance viewpoint. More generally, when Σ is not diagonal, $d_\Sigma(\underline{x}, \underline{\mu})$ defines a distance measure that takes into account the covariance (or correlation) among the variables, e.g., so that if 2 variables X_i and X_j were perfectly correlated then the contribution of the distances in dimensions x_i and x_j would be down-weighted since they are really measuring the same thing.

The Gaussian multivariate density function has the following properties:

- it has a single (unimodal) peak at $\underline{x} = \underline{\mu}$;
- the height of the density function decreases exponentially as a function of $d_\Sigma(\underline{x}, \underline{\mu})$, as \underline{x} moves further away from $\underline{\mu}$;
- the iso-contours of the density function (loci of points in \underline{x} space that all have the same value of $p(\underline{x})$) will be ellipsoidal in the general case (or “hyperellipsoidal” for $d > 2$). If Σ is diagonal, the major axes of the hyperellipse will be axis-parallel, and if $\Sigma = I$ then the isocontours are circles;
- any subset of X_i 's are also Gaussian, e.g., $p(X_1, X_3)$ is Gaussian as is $p(X_2)$, etc.

This will be covered at a later point—this is a placeholder for now.