Statistical Methods for Machine Learning II

© by Xia, Tingfeng

2020 winter term

Preface

This document is consist of notes from lectures and the online course notes that I, personally, find interesting/important. You can find the online course notes on the course website here: `https://probmlcourse.github.io/sta414/`

Contents

1 Lecture 2 - Introduction to Probabilistic Models

1.1 Overview

We have a random vector in the form $X = (X_1, \ldots, X_m)$ which can be ***either observed or unobserved***. To approach this in a generative way, we make the so called generative assumption. which is that $X \sim P_{true}(X)$, i.e. there is some true distribution that is behind the scene and our data is from such distribution.

Goal Model a parametric joint distribution $P_\theta(X)$ by learning the parameters. The learning here means we want to find a/the "close"/"best" estimation to our parameter θ. In this course we will investigate the following three problems,

- How to specify the joint, $P_\theta(X)$?
- What does "best"/"close" mean? In some sense we want to find $P_\theta \approx P_{true}$, however P_{true} might also be unknown.
- How to find the best θ? In this course we will generally rely on gradient methods, so ∇_θ...

1.2 Probabilistic Perspective on ML

With this perspective, we can think about common machine learning tasks differently, where random variables represent:

- X: (high dimensional) input data
- C: discrete label
- Y: continuous target

If we assume our knowledge of the joint of the above three, i.e. we know $P(X, C, Y)$, then we can write our familiar tasks in the following way

- ***Regression***:
$$p(Y|X) = \frac{p(X,Y)}{P(X)} = \frac{p(X,Y)}{\int p(X,Y)dY}$$
- ***Classification/Clustering***:
$$p(C|X) = \frac{p(X,C)}{\sum_{C'} p(X,C')}$$

1.2.1 (Example) Classification

Suppose we have data of the form $\mathcal{D} = \{(x, c)_i\}_i$. We assume that they came from a certain true distribution, i.e. $\{(x, c)_i\}_i \sim p(X, C)$. Then, the ultimate goal of the ML problem is converted into finding $p(C|X)$. Using Bayes Rule of total probability, we can expand the distribution of interest into

$$p(C|X) = \frac{p(X, C)}{P(X)} = \frac{p(X, C)}{\sum_{C'} p(X, C')}$$

Output Heuristics After we acquire $p(C|X)$ as above, we are one step away from our goal of output the actual prediction c^*. There are three ways that we can do this, namely

- ***MLE Estimate*** is the most intuitive one, we simply choose

$$c^* = \arg\max_c p(C = c|X)$$

- ***Sample Learnt Dist*** is another approach which produces non-deterministic results, i.e. we sample $c^* \sim p(C|X)$.

- ***Combined*** is usually a safe way of doing this. We output

$$(c^*, p(C = c^*|X)) \iff (\text{result}, \text{how sure?})$$

  As an example, we have a ML algorithm drives a car. In this case, we might want to make decision only when the machine learning model has a certain level of confidence.

1.3 Observed vs Unobserved Random Variables

1.3.1 Supervised Dataset

$$\{x_i, c_i\}_{i=1}^N \sim p(X, C)$$

In such case, the class labels are observed and finding the conditional distribution $p(C|X)$ satisfies the supervised classification problem.

1.3.2 Unsupervised Dataset

$$\{x_i\}_{i=1}^N \sim p(X, C)$$

Still under the generative assumption, where we assume that there is some underlying distribution for our dataset. Further, we assume that the distribution of data is related to the class labels for the data points even though the class labels are never observed. **A common way to refer to an unobserved discrete class label is "cluster"**. However, in this case, our final goal of classification is still $p(C|X)$[1].

[1] Might be helpful to think of Gaussian Mixture Models

1.3.3 Latent Variables

Further, like clusters, introducing assumptions about unobserved variables is a powerful modelling tool. We will make use of this by modelling variables which are never observed in the dataset, called latent or hidden variables. By introducing and modelling latent variables, we will be able to naturally describe and capture abstract features of our input data.

1.4 Operations on Probabilistic Models

- ***Generate Data***: Sample from the model.
- ***Estimate Likelihood***: When all variables are either observed or marginalized the result is a single real number which is the probability of the all variables taking on those specific values.
- ***Inference***: Compute the expected value of some variables given others which are either observed or marginalized.
- ***Learning***: Set the parameters of the joint distribution given some observed data to maximize the probability of the observed data.

1.5 Desiderata of Probabilistic Models

We have to desires for the joint distribution to learn, namely

- The marginal and conditional distribution can be computed efficiently
- The representation of the joint distribution should be compact. This is especially important when we are dealing with joint distributions over many variables.

In general, total joint distribution are too large to specify and would require an insane amount of data to fit even we wanted to. Thus, we need modelling assumptions.

1.5.1 Fully Dependent Factorization (Chain Rule)

Suppose we have sample space defined such that $|T| = 2$, $|W| = 3$, and $|M| = 4$. Then the total joint distribution could be expanding using the chain rule as (*note: not unique*)

$$P_\theta(T, W, M) = P(T)P(W|T)P(M|T, W)$$

which requires[2] $|\theta| = (2-1) + (3-1) \times 2 + (4-1) \times 2 \times 3 = 23$ parameters in total to specify.

[2]Here the bars mean "cardinality" rather than vector norm

1.5.2 Assumptions (Independence)

Introducing assumptions results in

- a less expressive model
- $|\theta|$ (usually, much) smaller

which can be bad sometimes and is a trade-off that we as modellers have to deal with.

(Example) Fully Independent We assume that $T \perp W \perp M$, then by definition we know, for example, $P(W|T) = P(W)$. Then

$$\begin{aligned} P_\theta(T, W, M) &= P(T)P(W|T)P(M|T, W) \\ &= P(T)P(W)P(M) \end{aligned}$$

which requires only $|\theta| = (2-1) + (3-1) + (4-1) = 6$ to fit.

1.6 Likelihood Function

For some observed data X, the likelihood describes the likeliness of the data under then distribution with parameter θ.

$$L(\theta) = p(X|\theta)$$

In general, we prefer to deal with the log likelihood function, which is defined as

$$\ell(\theta; X) = \log L(\theta) = \log p(x|\theta)$$

1.7 Maximum Likelihood Estimation

The idea is to find

$$\hat{\theta}_{MLE} := \arg\max_\theta \ell(\theta; \mathcal{D})$$

In the case of i.i.d, we can re-write as

$$\begin{aligned} \hat{\theta}_{MLE} &= \arg\max_\theta \ell(\theta; \mathcal{D}) \\ &= \arg\max_\theta \log \prod_m p\left(x^{(m)}|\theta\right) \\ &= \arg\max_\theta \sum_m \log p\left(x^{(m)}|\theta\right) \end{aligned}$$

1.8 Sufficient Statistics

(Definition) Statistic A statistic is a possibly vector valued *deterministic* function of a set of random variables.

(Definition) Sufficient Statistic is a statistic that conveys exactly the same information about the data generating process that created the data as the entire data itself.[3] In formal language, Sufficient Statistic ($T(X)$) for X could be defined as

$$T(X^{(1)}) = T(X^{(2)}) \implies L(\theta; X^1) = L(\theta; X^2), \quad \forall\theta$$

alternatively, we can define it as

$$P(\theta|T(X)) = P(\theta|X); \quad \text{i.e. data doesn't give further info}$$

1.8.1 Neyman Factorization

Equivalently to the above, we can write

$$P(\theta|T(X)) = h(x, T(x))g(T(x), \theta)$$

1.9 Exponential Family

Factorizes as

$$\begin{aligned} p(x|\eta) &= h(x)\exp\left\{\eta' T(x) - g(\eta)\right\} \\ &= h(x)g(\eta)\exp\left\{\eta' T(x)\right\} \end{aligned}$$

1.9.1 (Example) 1-D Gaussian

$$\begin{aligned} p(x|\theta) = \mathcal{N}(x|\mu,\sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x^2 - 2x\mu + \mu^2)\right) \\ &= \underbrace{\frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-\mu^2}{2\sigma^2}\right)}_{\log A(\eta)}\exp\left(\underbrace{\begin{bmatrix}\frac{\mu}{\sigma^2} & \frac{-1}{2\sigma^2}\end{bmatrix}}_{\eta}\underbrace{\begin{bmatrix}x \\ x^2\end{bmatrix}}_{T(X)}\right) \end{aligned}$$

2 Lecture 3 - Directed Graphical Models

2.1 Decision Theory (Utility Theory)

Let a denote action, and a^* be the optimal one. Also use s to denote state and $V(\cdot)$ be the value function. We have, in general

$$a^* = \arg\min_a / \arg\max_a \underbrace{\mathbb{E}_{p(s|a,knowledge)}\left[V(s)\right]}_{u(a) \triangleq \text{ utility of action } a}$$

[3] Could also be interpreted as "summarize the data with respect to the likelihood"

2.2 Graphical Model Notation

2.2.1 Chain Rule Expansion

Given any joint probability of N random variables, we can expand it as follows

$$p\left(x_{1,\ldots,N}\right) = p\left(x_1\right) p\left(x_2|x_1\right) p\left(x_3|x_2, x_1\right) \ldots p\left(x_n|x_{n-1:1}\right)$$

Formally speaking, in the case of two random variables would simply

$$p(x, y) = p(x|y)p(y)$$

and the general case for N random variables could be written as [4]

$$p\left(\bigcap_{i=1}^{N} x_i\right) = \prod_{j=1}^{N} p\left(x_j \middle| \bigcap_{k=1}^{j-1} x_k\right)$$

2.2.2 Graph Representation

(Example) Grouping Variables Consider the model

$$p\left(x_i, x_{\pi_i}\right) = p\left(x_{\pi_i}\right) p\left(x_i|x_{\pi_i}\right)$$

which we can use the following graph to represent:

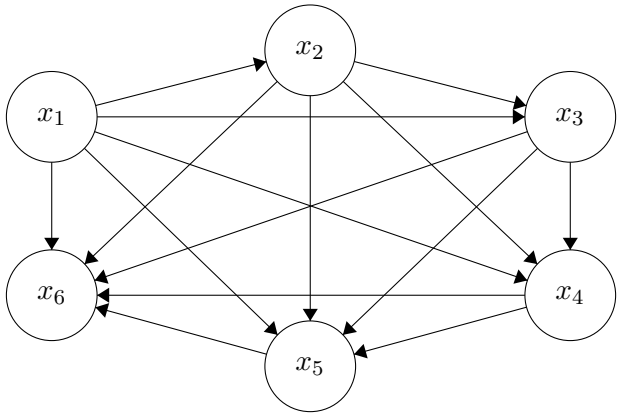
where

- ***nodes*** represent random variables
- ***arrows*** mean "conditioned on", e.g. "x_i is conditioned on x_{π_i}"

Notice that we can always group the variables together into one bigger variable, so in this example, x_{π_i} might represent a group of variables in stead of just one.

(Example) Fully Dependent 6 Nodes The total expansion of $p(x_{1:6})$ could be represented as

[4] Note: when $k = 1$, $p\left(x_k| \cap_{j=1}^{k-1} x_j\right) = p(x_1)$



This is the resultant graphical model if we make ***absolutely no assumption*** on the independence. Notice that such model grows exponentially in complexity with respect to the number of parameters considered. We say such model "scales poorly".

2.2.3 Conditional Independence

(Definition) Conditional Independence Let X be the set of nodes in our graph (the random variables of our model), then two sets of variables X_A, X_B are said to be conditionally independent given a third set of variables X_C if and only if either

$$p\left(X_A, X_B | X_C\right) = p\left(X_A | X_C\right) p\left(X_B | X_C\right)$$

or (***#! Important!***)

$$\begin{aligned} & p\left(X_A | X_B, X_C\right) = p\left(X_A | X_C\right) \\ \Longleftrightarrow\ & p\left(X_B | X_A, X_C\right) = p\left(X_B | X_C\right) \end{aligned}$$

and we denote the relation as $(X_A \perp X_B | X_C)$

2.2.4 Plates

In Bayesian methods, we treat parameters as random variables and hence we would like to include them in out graphical model. However, adding a node for each observation is quite cumbersome and thus we introduce plates, which denote replication of random variables.

Nested Plates Plates could be nested, in which case their arrows get duplicated also, ***according to the rule:*** draw an arrow from every copy of the source node to every copy of the destination node.

Crossing Plates Plates can also cross (intersect), in which case the nodes at the intersection have multiple indices and get duplicated a number of times equal to the product of the duplication numbers on all the plates containing them.

2.3 Directed Acyclic Graphical Models (DAGM)

A directed acyclic graphical model over N random variables look like

$$p\left(x_{1:N}\right) = \prod_{i}^{N} p\left(x_i | x_{\pi_i}\right)$$

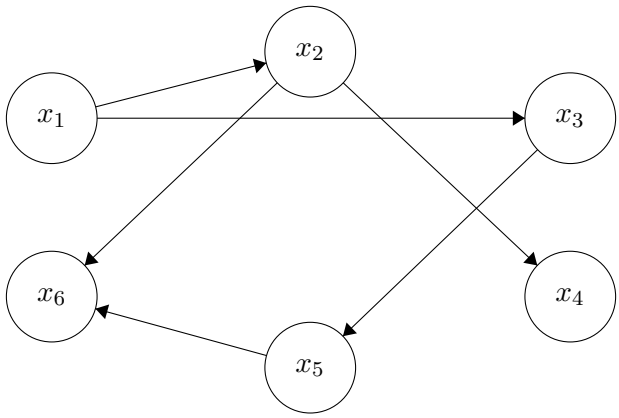
where x_i is a random variable and x_{π_i} denotes the parents of the node (which could be an empty set). This notion is more general than the fully dependence model that looked at above. Notice that here each node is only dependent on its parents rather than all other nodes. Thus, the complexity of such model reduces to exponential in fan-in of each node, instead of the total N.

2.3.1 Independence Assumption on DAGMs

Then, we have the independence relationship of[5] $x_i \perp x_{Ancestor(\pi_i)} | x_{\pi_i}$ which expands into

$$p\left(x_{1,\ldots,6}\right) = p\left(x_1\right) p\left(x_2|x_1\right) p\left(x_3|x_1\right) p\left(x_4|x_2\right) p\left(x_5|x_3\right) p\left(x_6|x_2, x_5\right)$$

with the following respective graph



As we can see, the introduction of the assumption greatly reduced the complexity of the model.

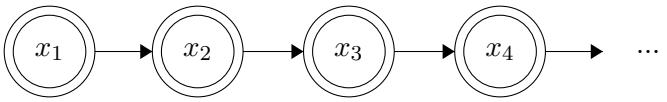
[5] Requires topological ordering, to be added later.

2.3.2 (Example) Markov Chain

The following example has independence relationships that satisfies the Markov Property.

$$p(x) = p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_3)...$$

and could be represented, in graphical model, as



2.4 Directed - Separation

2.4.1 (Definition) D - Separation

Directed-separation is a notion of connectedness in DAGs in which two (sets of) variables may or may not be connected conditioned on a third (set of) variable(s). D-connection implies conditional dependence and d-separation implies conditional independence.

2.4.2 Float Rules Derivation

Notation The double circles (or shaded circles) represent the notion of "conditioned-on".

Chain

In this case, we are interested in knowing whether or not

$$(X \perp Z)|Y$$

From the arrows between the nodes, we know that

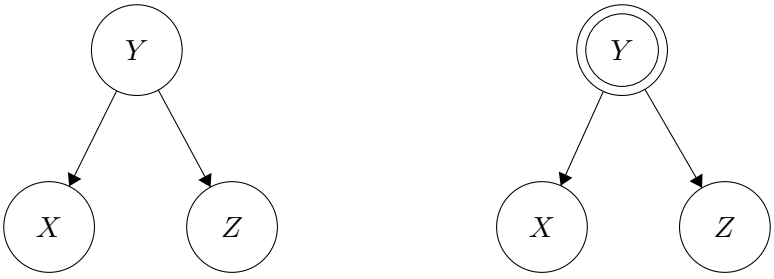
$$P(X, Y, Z) = P(X)P(Y|X)P(Z|Y)$$

Then, we have

$$\begin{aligned} P(X, Z|Y) &= \frac{P(X, Y, Z)}{P(Y)} \\ &= \frac{P(X)P(Y|X)P(Z|Y)}{P(Y)} \\ &= \frac{P(X, Y)P(Z|Y)}{P(Y)} \\ &= P(X|Y)P(Z|Y) \end{aligned}$$

and this completes the proof. ■

Common Clause



As previous, we are interested in knowing whether or not $(X \perp Z)|Y$. From the graph, we can know that

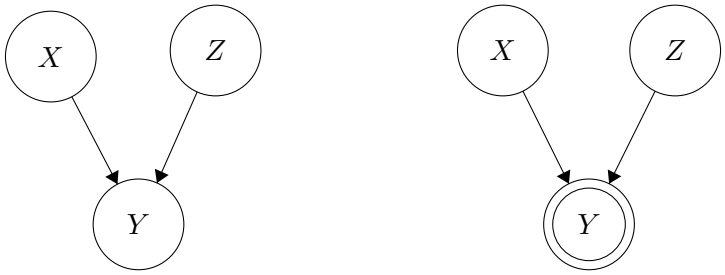
$$P(X, Y, Z) = P(Y)P(X|Y)P(Z|Y)$$

Then, we have

$$\begin{aligned} P(X, Z|Y) &= \frac{P(X, Y, Z)}{P(Y)} \\ &= \frac{P(Y)P(X|Y)P(Z|Y)}{P(Y)} \\ &= P(X|Y)P(Z|Y) \end{aligned}$$

and this completes the proof. ■

Explaining Away (Berkson's Paradox)



Notice from the graph that we have $P(X, Y, Z) = P(X)P(Z)P(Y|X, Z)$. First, I will prove that, in this case, $(X \not\perp Z)|Y$.

$$\begin{aligned} P(Z|X, Y) &= \frac{P(X)P(Z)P(Y|X, Z)}{P(X)P(Y|X)} \\ &= \frac{P(Z)P(Y|X, Z)}{P(Y|X)} \\ &\neq P(Z|Y) \end{aligned}$$

For marginal independence, i.e. $(X|Z)$, we want to show that $P(X,Z) = P(X)P(Z)$.

$$\begin{aligned} P(X,Z) &= \sum_{Y'} P(X,Y',Z) \\ &= \sum_{Y'} P(X)P(Z)P(Y'|X,Z) \\ &= P(X)P(Z) \end{aligned}$$

■

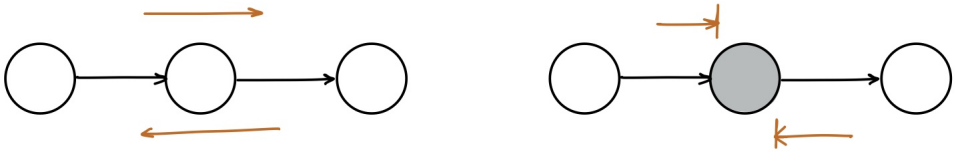
2.4.3 The Bayes-Ball Algorithm

To check if $x_A \perp x_B | x_C$, we will need

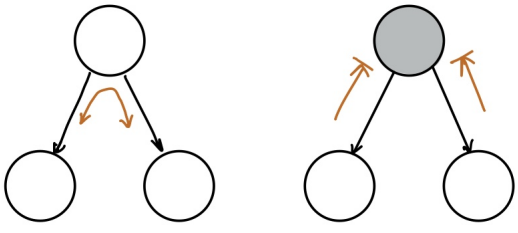
- Shade all nodes that were conditioned on, i.e. all nodes of x_C
- Place balls at each node in x_A (or x_B)
- Let the balls bounce around according to rules that are yet to be states below
  - If any of the balls reach any of the nodes in x_B from x_A (or reach x_A from x_B) then we declare $x_A \not\perp x_B | x_C$
  - Otherwise, $x_A \perp x_B | x_C$

2.4.4 Bayes Ball Rules

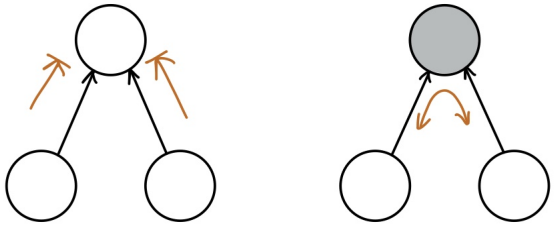
Chain



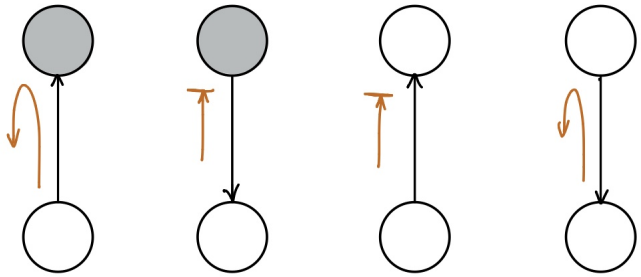
Common Cause



Explain Away



Linear



2.5 Unobserved Variables

Certain variables in our models may be unobserved , either some of the time or always, at training time or at test time. Graphically, we use shading to indicate observation.

2.5.1 Partially Unobserved Variables

[6] If variables are occasionally unobserved hen they are missing data, e.g., undefined inputs, missing class labels, erroneous target values. In this case, we can still model the joint distribution, but we marginalize the missing values

$$\begin{aligned}\ell(\theta;\mathcal{D}) &= \sum_{\text{complete}} \log p\left(x^c, y^c|\theta\right) + \sum_{\text{missing}} \log p\left(x^m|\theta\right) \\ &= \sum_{\text{complete}} \log p\left(x^c, y^c|\theta\right) + \sum_{\text{missing}} \log \sum_y p\left(x^m, y|\theta\right)\end{aligned}$$

2.5.2 Latent Variables

Above we discussed the case where some data are non-deterministically unobserved. Latent variables refers to those that ***never observed***. The handling of the latent variables depends on where it appears in our model,

[6] An concrete example would be hospital data, where typically a large proportion of the data is missing.

- If we never condition on it when computing the probability of the variables we do observe, then we can just forget about it and integrate it out. For example, given y, x we fit the model
$$p(z, y|x) = p(z|y)p(y|x, w)p(w)$$
- If z is not a leaf node, marginalizing over it will induce dependencies between its children. For example, given y, x we can fit the model
$$p(y|x) = \sum_z p(y|x, z)p(z)$$

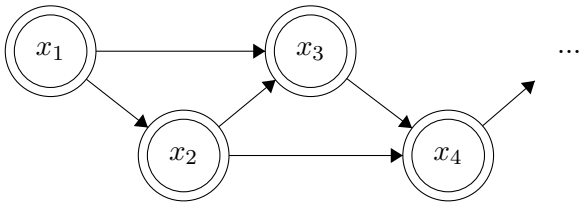
2.5.3 Mixture Models

In this case, we are looking at data that has no input on class information. We can sum the labels out,
$$p(x|\theta) = \sum_{k=1}^{K} p\left(z = k|\theta_z\right) p\left(x|z = k, \theta_k\right)$$
where bayes rule comes in handy for calculating the posterior (class responsibilities) of the mixture component given some data, i.e.,
$$p\left(z = k|x, \theta_z\right) = \frac{p\left(z = k|\theta_z\right) p_k\left(x|\theta_k\right)}{\sum_j p\left(z = j|\theta_z\right) p_j\left(x|\theta_j\right)}$$

2.6 Examples

2.6.1 Second-order Markov Chain

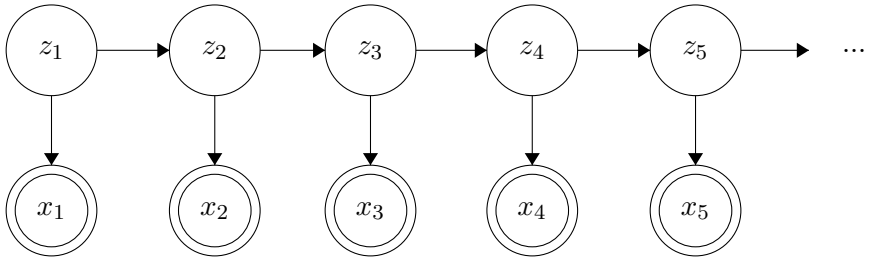
Consider the model
$$p\left(\mathbf{x}_{1:T}\right) = p\left(x_1, x_2\right) p\left(x_3|x_1, x_2\right) p\left(x_4|x_2, x_3\right) \cdots = p\left(x_1, x_2\right) \prod_{t=3}^{T} p\left(x_t|x_{t-1}, x_{t-2}\right)$$
which has the graphical representation (double circles mean "observed" here)



we notice that this model essentially assumes "the present depends on the past only through the current state as well as the last one."

2.6.2 Hidden Markov Models (HMMs)

HMM is a statistical model in which a system being modelled is assumed to be a Markov Process[7] with un-observed states. Here is the graphical model, where double circles mean "observed" here:



where

- z_t are hidden states taking on one of K discrete values
- x_t are observed variables taking on values in any space.

The above graph factorizes into

$$p\left(X_{1:T}, Z_{1:T}\right) = p\left(Z_{1:T}\right) p\left(X_{1:T} | Z_{1:T}\right) = p\left(Z_1\right) \prod_{t=2}^{T} p\left(Z_t | Z_{t-1}\right) \prod_{t=1}^{T} p\left(X_t | Z_t\right)$$

[7] In continuous time, the Markov Chain model is known as Markov Process