STA 414/2104:
Statistical Methods in Machine Learning II

Week 2 : Decision Theory & Directed Graphical Models

Piotr Zwiernik

University of Toronto

Table of contents

1. Statistical decision theory

   Optimal decisions in classification

   Optimal decisions for regression

2. Directed graphical models

   Basic definitions

   Conditional independence on Directed Acyclic Graphs (DAGs)

   Bayes Ball Algorithm

Statistical decision theory

Decision making

Framework for understanding many of the procedures we consider.

- Suppose we have an input vector x and a corresponding target (output) value c with joint probability distribution: $p(x, c)$.
- Our goal is to predict the output label c given a new value for x.
- For now, we focus on classification so c is a categorical variable, but the same reasoning applies to regression (continuous target).

The joint probability distribution $p(x, c)$ provides a complete summary of uncertainties associated with these random variables.

Example: Cancer screening from chest X-ray

Based on the X-ray image, we would like determine if the patient has cancer or not.

- The input vector x is pixel intensities, and the output c represents the presence of cancer, class $\mathcal{C}_1$, or absence of cancer, class $\mathcal{C}_2$.

- $\mathcal{C}_1$ cancer present
- $\mathcal{C}_2$ cancer absent

We can use an "arbitrary" encoding for these classes $\mathcal{C}_1$ and $\mathcal{C}_2$, e.g. take: $c = 0$ correspond to class $\mathcal{C}_1$, and $c = 1$ corresponds to $\mathcal{C}_2$.

Optimal decisions

Decision Problem

Suppose we estimated the joint distribution $p(x, c)$ using some ML method. Decide whether to give treatment to the patient or not.

- Given a new X-ray image, our goal is to decide which of the two classes that image should be assigned to. We could compute conditional probabilities of the two classes, given the input image, for $k = 1, 2$:

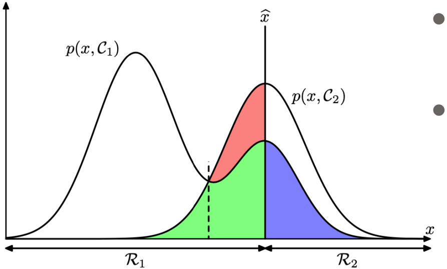
$$p(\mathcal{C}_k|x) = \frac{p(x|\mathcal{C}_k)p(\mathcal{C}_k)}{p(x)} \quad \text{Bayes' rule.}$$

- Intuitively, pick class with higher posterior probability.
- We now formalize in what sense this choice is optimal.

Misclassification rate

Decision rule: Divide the input space into regions $\mathcal{R}_1, \mathcal{R}_2$ (decision regions) such that all points in $\mathcal{R}_k$ are assigned to class $\mathcal{C}_k$, $k = 1, 2$.

Criterion to optimize: Make as few misclassifications as possible.

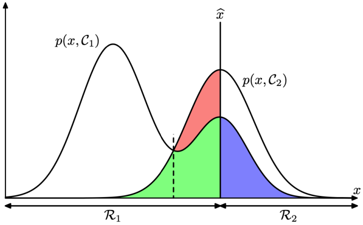


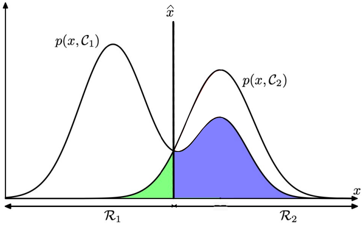
- Red + green regions: input belongs to class $\mathcal{C}_2$, but is assigned to $\mathcal{C}_1$.
- Blue region: input belongs to class $\mathcal{C}_1$, but is assigned to $\mathcal{C}_2$.

$$p(\text{mistake}) \;=\; p(x \in \mathcal{R}_1, \mathcal{C}_2) + p(x \in \mathcal{R}_2, \mathcal{C}_1) \;=\; \int_{\mathcal{R}_1} p(x, \mathcal{C}_2) dx + \int_{\mathcal{R}_2} p(x, \mathcal{C}_1) dx$$

Misclassification rate

Compare the following two decision rules:





- Blue + green area is always included in the p(mistake).
- On the left there are points $x \in \mathcal{R}_1$ for which $p(x, \mathcal{C}_2) > p(x, \mathcal{C}_1)$ (red part)
- Reduce the red area by moving the threshold $\hat{x}$ to the left.

Misclassification error

- Misclassification error:

$$p(\text{mistake}) = \underbrace{\int_{\mathcal{R}_1} p(x, \mathcal{C}_2) dx}_{\text{red+green}} + \underbrace{\int_{\mathcal{R}_2} p(x, \mathcal{C}_1) dx}_{\text{blue}}$$

  and the decision regions $\mathcal{R}_1$ and $\mathcal{R}_2$ are disjoint.
- Therefore, for a particular input x, if $p(x, \mathcal{C}_1) > p(x, \mathcal{C}_2)$, then we assign x to class $\mathcal{C}_1$. I.e. $\mathcal{R}_1 = \{x : p(x, \mathcal{C}_1) > p(x, \mathcal{C}_2)\}$.

Minimizing misclassification

Since $p(x, \mathcal{C}_k) = p(\mathcal{C}_k|x)p(x)$, in order to minimize the probability of making mistake, we assign each x to the class for which the posterior probability $p(\mathcal{C}_k|x)$ is largest. This minimizes the misclassification rate.

Expected loss

Simply minimizing the missclassification rate may not be desirable.

- We incorporate a **loss function** to measure the loss incurred by taking any of the available decisions.
- Suppose that for x, the true class is $\mathcal{C}_k$, but we assign x to class $\mathcal{C}_j$ and incur loss of L_{kj} ((k, j)-th element of a loss matrix).

Consider medical diagnosis example: example of a loss matrix:

$$\begin{array}{cc} & \textbf{Decision} \\ & \begin{array}{cc}\text{cancer} & \text{normal}\end{array} \\ \textbf{Truth}\ \begin{array}{c}\text{cancer}\\ \text{normal}\end{array} & \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix}\end{array}$$

Incorrectly classify as healthy

Incorrectly classify as cancer

Thus the expected loss is given by

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj}\ p(x, \mathcal{C}_k) dx$$

New goal:Minimize expected loss

Therefore, we want to minimize

$$\begin{aligned}\mathbb{E}[L] &= \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj}\; p(x, \mathcal{C}_k) dx \\ &= \sum_j \int_{\mathcal{R}_j} \sum_k L_{kj}\; p(x, \mathcal{C}_k) dx.\end{aligned}$$

Define $g_j(x) = \sum_k L_{kj}\; p(x, \mathcal{C}_k)$. Notice that $g_j(x) \geq 0$ and

$$\mathbb{E}[L] = \sum_j \int_{\mathcal{R}_j} g_j(x) dx$$

Thus, minimizing $\mathbb{E}[L]$ is equivalent to choosing

$$\mathcal{R}_j = \{x \;:\; g_j(x) < g_i(x) \;\text{ for all }\; i \neq j\}.$$

Simplifying further

We can also use the product rule $p(x, \mathcal{C}_1) = p(\mathcal{C}_1|x)p(x)$ and reduce the problem to:

Discriminant rules:

Find regions $\mathcal{R}_j$ such that the following is minimized:

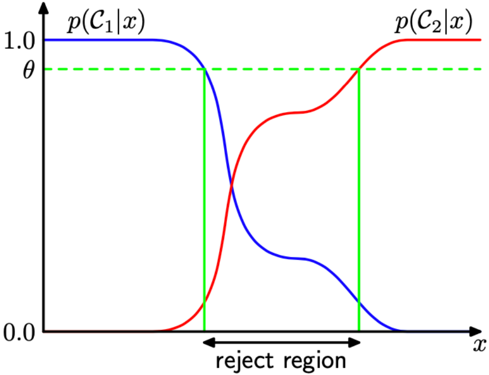
$$\sum_k L_{kj} p(\mathcal{C}_k|x).$$

That is

$$\mathcal{R}_j = \left\{x \ : \sum_k L_{kj} \ p(\mathcal{C}_k|x) < \sum_k L_{ki} \ p(\mathcal{C}_k|x) \ \text{ for all } \ i \neq j\right\}.$$

Reject option

For the regions where we are relatively uncertain about class membership, we do not have to make a decision.



When the conditional class probabilities fall below θ, we refuse to make a decision.

Loss functions for regression

- Consider an input/target setup (x, t) where the target (output) is continuous $t \in \mathbb{R}$, and the joint density is $p(x, t)$.
- We aim to find a regression function $y(x) \approx t$ which maps inputs to the outputs.
- Consider the squared loss function L between $y(x)$ and t to assess the quality of our estimate $L(y(x), t) = (y(x) - t)^2$.

Goal:

What is the best function $y(x)$ that minimizes the expected loss?

$$\mathbb{E}[L] = \int \int L(y(x), t) p(x, t) dxdt.$$

Minimizing expected loss: Best regression function

We add and subtract $\mathbb{E}[t|x]$ and write

$$\begin{aligned}
\mathbb{E}[L] &= \int\int (y(x) - t)^2 p(x,t)dxdt \\
&= \int\int (y(x) - \mathbb{E}[t|x] + \mathbb{E}[t|x] - t)^2 p(x,t)dxdt \\
&= \int\int (y(x) - \mathbb{E}[t|x])^2 p(x,t)dxdt + \int\int (\mathbb{E}[t|x] - t)^2 p(x,t)dxdt \\
&\quad + 2\int\int (y(x) - \mathbb{E}[t|x])(\mathbb{E}[t|x] - t)p(x,t)dxdt
\end{aligned}$$

The last term is zero since

$$\begin{aligned}
&\int\int (y(x) - \mathbb{E}[t|x])(\mathbb{E}[t|x] - t)p(x,t)dxdt \\
&= \int\int (y(x) - \mathbb{E}[t|x])(\mathbb{E}[t|x] - t)p(t|x)p(x)dxdt \\
&= \int (y(x) - \mathbb{E}[t|x])\Big\{\underbrace{\int (\mathbb{E}[t|x] - t)p(t|x)dt}_{=0}\Big\}p(x)dx = 0
\end{aligned}$$

Best regression function

- We showed that the expected loss is given by the sum of two **non-negative** terms

$$\mathbb{E}[L] = \int\int (y(x) - \mathbb{E}[t|x])^2 p(x,t)dxdt + \int\int (\mathbb{E}[t|x] - t)^2 p(x,t)dxdt.$$

- The second term does not depend on $y(x)$ thus choosing the best regression function $y(x)$ is equivalent to minimizing the first term on the right hand side.
- This term is always non-negative and exactly zero if

$$y(x) = \mathbb{E}[t|x].$$

- The second term is the expectation of the conditional variance of $t|x$. It represents the intrinsic variability of the target data and can be regarded as noise.

Summary: Decision making

- Depending on the application, one needs to choose an appropriate loss function.
- Loss function can significantly change the optimal decision rule.
- One can always use the reject option and not make a decision.
- In case of regression, the optimal map between x and t corresponds to the conditional expectation $\mathbb{E}[t|x]$.
- We focuse on classification/regression but similar framework can be used to evaluate any statistical procedure (e.g. estimation).

Directed graphical models

Next:

- Graphical models notation
- Conditional independence on directed acyclic graphs (DAGs)
- Bayes Ball

Joint distributions

- The joint distribution of N random variables $(x_1, x_2, ..., x_N)$ is a very general way to encode knowledge about a system.
- Assume $x_i \in \{0, 1\}$ are binary, then it requires $2^N - 1$ parameters to specify the joint distribution

$$p(x_1, x_2, ..., x_N).$$

- This can be also written as

$$p(x_1, x_2, \ldots, x_N) = \prod_{j=1}^{N} p(x_j | x_1, x_2, \ldots, x_{j-1})$$

for any ordering of the variables, where $p(x_1|x_0) = p(x_1)$.

Powerful modelling principle

Exploit dependencies among variables and reduce the number of parameters!

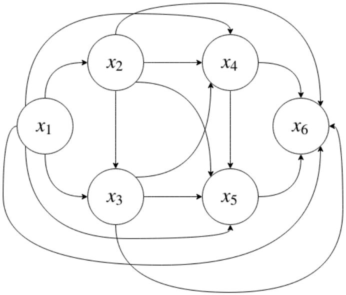
Conditional Independence

- Assume there are N random variables $x_1, x_2, ..., x_N$.
- For set $A \subset \{1, 2, ..., N\}$, we denote by $x_A = \{x_i \ : \ i \in A\}$.
- For disjoint A, B, C, if random variables x_A, x_B are conditionally independent given x_C, we write

$$x_A \perp x_B \mid x_C$$

- The following conditions are equivalent
  - $x_A \perp x_B | x_C$
  - $p(x_A, x_B | x_C) = p(x_A | x_C) p(x_B | x_C)$
  - $p(x_A | x_B, x_C) = p(x_A | x_C)$
  - $p(x_B | x_A, x_C) = p(x_B | x_C)$

Directed Acyclic Graphical Models (Bayes' Nets)



- A directed acyclic graphical model (DAG) encodes a particular form of factorization of the joint distribution.
- Variables are represented by nodes, and edges represent direct dependence.

DAG induces the following factorization of the joint distribution:

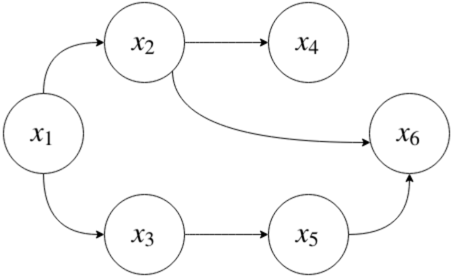
$$p(x_1, \dots, x_N) = \prod_{i=1}^{N} p(x_i | x_1, \dots, x_{i-1}) = \prod_{i=1}^{N} p(x_i | \text{parents}(x_i))$$

where parents(x_i) is the set of nodes with edges pointing to x_i.

Example: Joint factorization induced by a DAG

Recall: In a DAGs $p(x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} p(x_i | \text{parents}(x_i))$.

Consider the following graph:

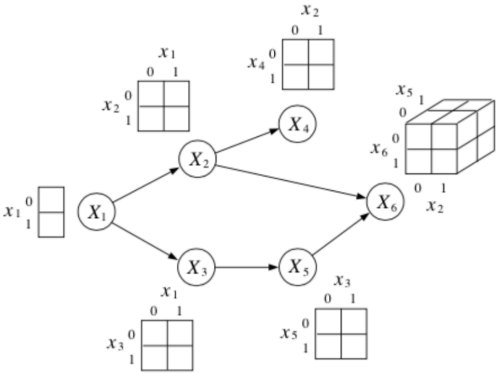


It induces the following factorization of the joint distribution:

$$p(x_1, x_2, ..., x_6) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2, x_5)$$

Conditional Probability Tables (CPT)

In our example, suppose each x_i is a binary random variable. How many parameters does it take to represent this joint distribution?



- For example, 2x2 CPT for the node x_4 corresponds to $p(x_4|x_2)$ requires 2 parameters.
- Each CPT with K_i parents requires 2^{K_i} parameters. In total, $\sum_i 2^{K_i} \leq N2^{\max K_i}$ parameters.
- If we allow all possible dependencies (fully-connected DAG), we need $2^N - 1$ parameters.

This gives a big reduction in storage and computations; here 63 vs 13.

Conditional Independence in DAGs

D-separation (directed-separation) is a notion of connectedness in DAGs in which two sets of nodes may or may not be connected conditioned on a third set of nodes.

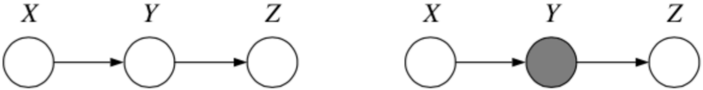
- Fix a DAG over N nodes $1, 2, \ldots, N$.
- This DAG defined factorization of the joint distribution $p(x_1, \ldots, x_N)$.
- This factorization implies some conditional independence that can be deducted from d-separation: if C d-separates A and B in the DAG then $x_A \perp x_B | x_C$.

We still have not defined d-separation...

Important reduction

- We have $x_A \perp x_B | x_C$ if and only if $x_a \perp x_b | x_C$ for all $a \in A$, $b \in B$.
- Also C d-separates A and B if and only if it d-separates each $a \in A$ and $b \in B$.

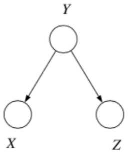
Causal Chain

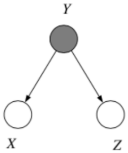


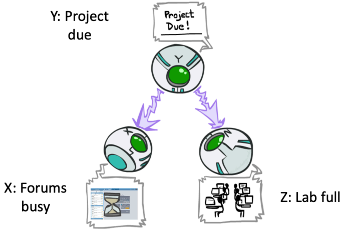
$$
\begin{aligned}
p(z|x,y) &= \frac{p(x,y,z)}{p(x,y)} \\
&= \frac{p(x)p(y|x)p(z|y)}{p(x)p(y|x)} \\
&= p(z|y) \quad X \text{ and } Z \text{ d-separated given } Y.
\end{aligned}
$$

Common Cause

Where we think of y as the "common cause" of the two independent effects x and z.







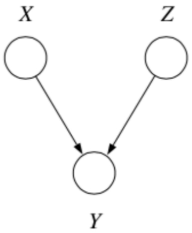
Question: When we condition on y, are x and z independent?

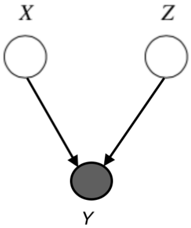
Answer: From the graph, we get

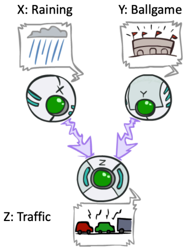
$$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(y)p(x|y)p(z|y)}{p(y)} = p(x|y)p(z|y) \quad \text{yes!}$$

Thus, Y d-seperates X and Z like in the previous case.

Explaining Away (Common Effect)







Question: When we condition on y, are x and z independent?

Answer: From the graph, we get

$$\begin{aligned} p(z|x,y) &= \frac{p(x)p(z)p(y|x,z)}{p(x)p(y|x)} \\ &= \frac{p(z)p(y|x,z)}{p(y|x)} \neq p(z|y) \end{aligned}$$

images credit Abbeel & Klein

Bayes Ball Algorithm

The Bayes Ball algorithm determes conditional independence in a DAG.

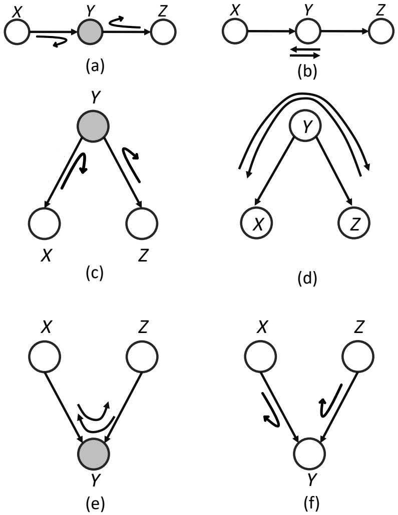
- To check if $x_A \perp x_B | x_C$ we need to check if every variable in A is d-seperated from every variable in B conditioned on all variables in C.

In general, the algorithm works as follows:

1. Shade all nodes x_C (these are observed)
2. Place "balls" at each node in x_A (or x_B)
3. Let the "balls" "bounce" around according to some rules
   - If any of the balls reach any of the nodes in x_B from x_A then $x_A \not\perp x_B | x_C$
   - Otherwise $x_A \perp x_B | x_C$

If any of the vertices in x_A are connected by an edge to a vertex in x_B, no conditional independence possible.

Bayes Ball: Rules for active/inactive triples

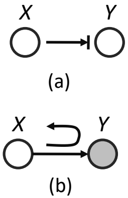


(a) (b) (c) (d) (e) (f)

Move from X to Z (or Z to X) crossing Y...

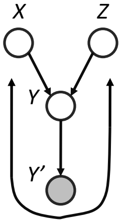
- Arrows: paths the balls can travel
- Arrows with bars: paths the balls cannot travel
- Notice balls can travel opposite to edge directions!

Boundary cases (Y is a leaf):



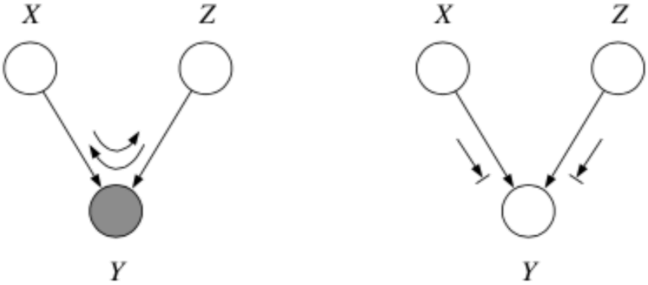
(a) (b)

One motivating example:



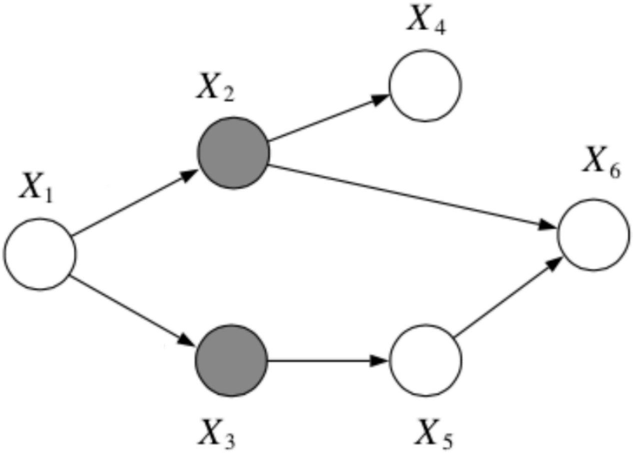
Example I: Explaining Away

If y or any of its descendants is shaded, the ball passes through.



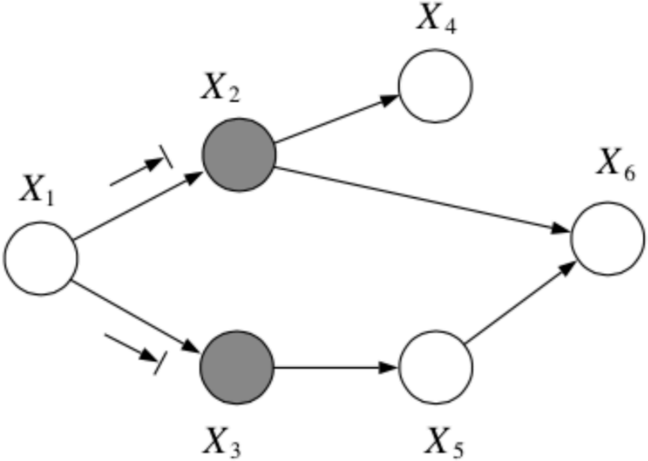
Example II: Large DAG

In the following graph, is $x_1 \perp x_6 | \{x_2, x_3\}$?



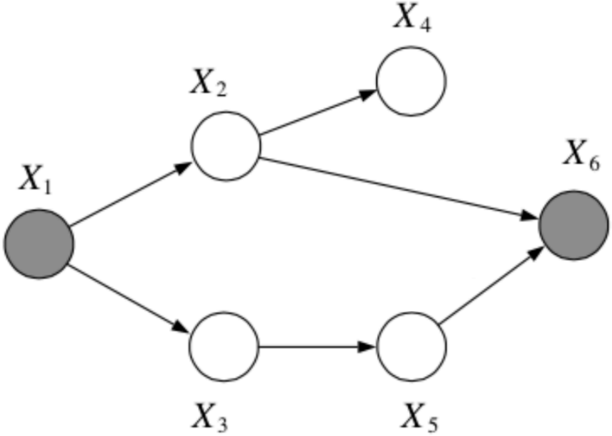
Example II: Solution

Yes, by the Bayes Ball algorithm.



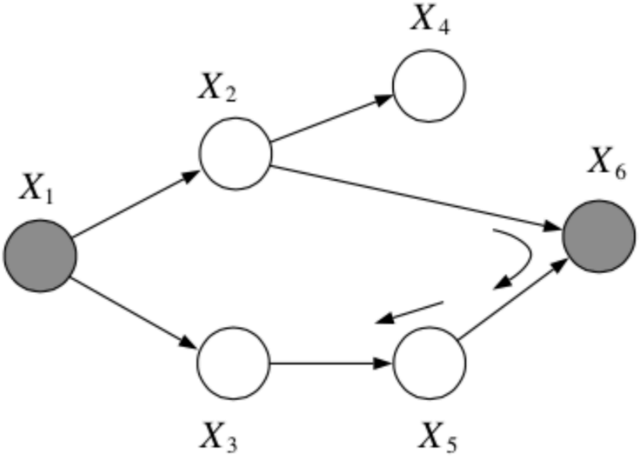
Example III:

In the following graph, is $x_2 \perp x_3 | \{x_1, x_6\}$?



Example III:

No, by the Bayes Balls algorithm.



Summary

- DAGs are great for encoding conditional independencies.
- They can reduce the number of parameters significantly.
- Conditional independence between two sets of variables on a DAG can be found using the Bayes ball method.
- Next lecture: Markov Random Fields.