ECON484 Machine Learning

6. Decision Trees (Part 1, Prerequisites)

Lecturer:
Bora GÜNGÖREN
bora.gungoren@atilim.edu.tr

Bayesian Decision Theory

- Data comes from a process of data acquisition which is not perfect. Therefore, data comes from a process that involves **unknown knowledge** as well.
    - We **choose** to model systems with random variables to account for this unknown data.
        - The systems we are trying to model could as well be deterministic, however due to our lack of knowledge we are forced to make that choice.
        - If we had access to that unknown knowledge, maybe we could model as deterministic.
    - The knowledge we have no access to is termed as **unobservable variables**.
        - If we denote the **observable variables** as x, and the unobservable variables as z we would end up with a relationship in the form of **x = f(z)** in our models. That is we observe the observables, and they are a result of the unobservables. However we can not come up with a practical model out of this relationship.
        - Therefore we choose to say **X is a random variable**, that is from the distribution **P (X = x)** that specifies the observable variables.
        - As usual, we do not know P(X) so we try to **estimate it from a given sample**.

Bayesian Decision Theory

- Classification problems are easy to model in this fashion.
- Suppose We are given a problem to do a binary classification, for example trying to classify an automobile as “safe” or “not safe”.
  - Let x_1 and x_2 be two observable variables (ie. chassis strength, weight). We choose to observe them because we have an intuition that they have something to do with the safety of a vehicle. Then we can define random variables X_1 and X_2 accordingly.
  - We can also define S as the safety outcome. Let S=0 denote safe and S=1 denote note safe.
  - Then the conditional probability $P(S \mid X_1, X_2)$ is the basis for our decisions.
    - Choose the vehicle is safe id the probability of it being safe given particular x_1,x_2 is higher than the probability of being unsafe given same particular x_1, x_2.
    - Choose S=0 if $P(S = 0 \mid X_1 = x_1, X_2 = x_2) > P(S = 1 \mid X_1 = x_1, X_2 = x_2)$

Bayesian Decision Theory

- Our method
  - Choose S=0 if $P(S = 0 \mid X_1 = x_1, X_2 = x_2) > P(S = 1 \mid X_1 = x_1, X_2 = x_2)$
  - Bayes’ rule gives us a chance to express this conditional probability as
    - $P(S \mid x) = P(S)\, P(x \mid S) / P(x)$
    - $P(S = 0 \mid x) = P(S=0)\, P(x \mid S = 0) / P(x)$
      - $P(S = 0)$ is called the **prior probability** and can be calculated from past samples.
        - The prior probability has no error, because it is based on past observations. Therefore $P(S = 0) + P(S = 1) = 1$.
      - $P(x \mid S = 0)$ is called the **class likelihood**, and can be calculated from past samples.
      - $P(x)$ is called the **evidence**, and represents the marginal probability that x is observed regardless of the case that the vehicle is safe or not safe. It can also be calculated from past samples.
    - Posterior = Prior x Likelihood / Evidence
      - The posteriors sum up to 1. **$P(S=0|x) + P(S=1|x) = 1$**.
      - Once we have the posteriors calculated, we go back to
        - Choose S=0 if $P(S = 0 \mid x) > P(S = 1 \mid x)$

Bayesian Decision Theory

- Our method can be extended to n-ary classification
  - We can extend any n-ary classification as a series of binary classifications as long as the probabilities of all classes are **mutually exclusive** and **exhaustive**.
    - Let's denote S=i as S_i
    - $1 \geq P(S_i) \geq 0$ and $\Sigma\ P(S_i) = 1$
  - How does the formulation change?
    - $P\ (S_i \mid x) = P\ (S_i)\ P\ (x \mid S_i)\ /\ P\ (x)$
    - $P(x) = \Sigma\ P(x|S_i)\ P(S_i)$
    - Choose S_i if $P\ (S_i \mid x) = \max P(S_i \mid x)$

Bayesian Decision Theory

- Once we calculate posterior probabilities, each classification instance (ie. trial) becomes deterministic in itself.
  - Our model includes no error term because of our **<u>assumption</u>** that the class probabilities are mutually exclusive and **<u>exhaustive</u>**. In a sense we **<u>overload the error term into one or more of the classes</u>**.
  - However, in reality we will have error.
  - This error is realized by two concepts: loss and risk.
    - When we classify a vehicle as S=0 (safe), we might be correct. That will reduce the **risk** we are taking. If we were incorrect (the vehicle was not safe) then we increased the risk we are taking.
    - When we classify a vehicle as S=1 (unsafe), and we were correct, we reduced the **loss** we would be facing. If we were incorrect (misclassified a safe vehicle as unsafe) then we increased the loss.

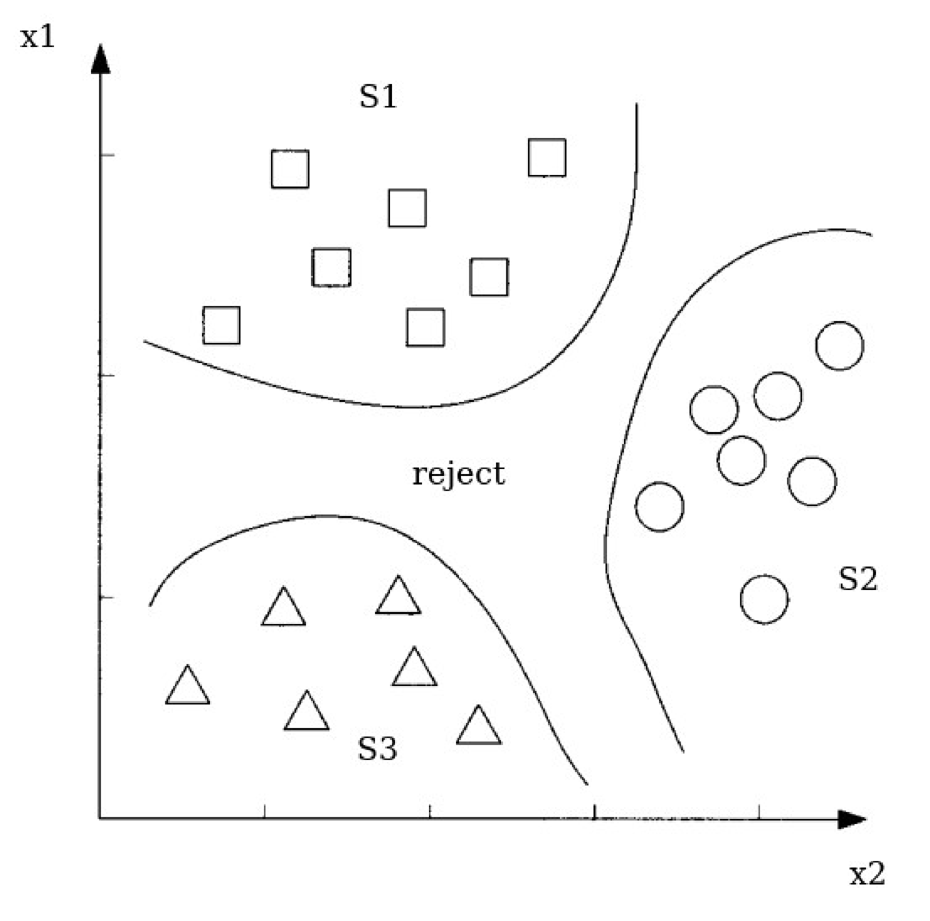
Bayesian Decision Theory

- Functions for loss and risk can be defined.
  - Let's define α_i as the decision to choose S=i (based on our estimated posterior probabilities).
  - Let's also define λ_{ik} as the loss due to mistakenly choosing S=i when it was actually S=k.
  - Then the expected risk of the action (decision) α_i is defined as
    - $R(\alpha_i|x) = \Sigma\ \lambda_{ik} P(C_k|x)$
    - so that we choose based on **risk minimization**
    - Choose α_i (ie. S=i) if $R(\alpha_i|x) = \min R(\alpha_k|x)$
  - So the decision is now a function of the losses, λ_{ik} which is apparently in a matrix form.
    - A special case is when i=k the loss is zero, and equal to 1 in all other cases. This is called a zero-one loss.
      - In this case the calculations are simplified so that $R(\alpha_i|x) = 1 - P(S_i|x)$
      - We choose the most probable case to minimize risk.
      - This is a very rare case, and far from realistic. However human intuitive decision making often assumes equal loss.

Bayesian Decision Theory

- If wrong decisions have a very high cost in your application area (ie. finance, health) we add **a new class called doubt**.
  - In practice when we assign to doubt (or reject) class, a **human operator** should review the case.
  - In the mathematical representation, i=1,..,K represents the K classes, and i=K+1 represents the doubt/reject class.
    - In this model λ_{ik} has the value 0 for i=k, and 1 for all other classes, except for the doubt/reject class. For i=K+1, we set $\lambda_{ik}=\lambda$ as a separate value and $0 < \lambda < 1$.
    - The risk of doubt/reject can be calculated just as risk of choosing any class.
    - Our rule is to mark as doubt/reject if the risk of reject is less than the risk of any other class. If this is not the case, then we assign to the class with minimum risk.
    - If $\lambda=0$ then we will always reject. If $\lambda \geq 1$ we will never reject.
- Classification can also be seen as implementing **a set of discriminant functions**.
  - In this case maximizing the discriminant function is equivalent to minimizing the risk function.

Bayesian Decision Theory



Bayesian Decision Theory

- Utility Theory generalizes the approach in risk minimization.
  - U_{ik} is the utility of action taking action i (α_i) when the state is at k (S_k).
  - So the expected utility is $EU(\alpha_i|x) = \Sigma\ U_{ik}P(S_k|x)$
  - We make our choices in order to maximize expected utility.
- In the limited context of classification, utility maximization corresponds to risk minimization.

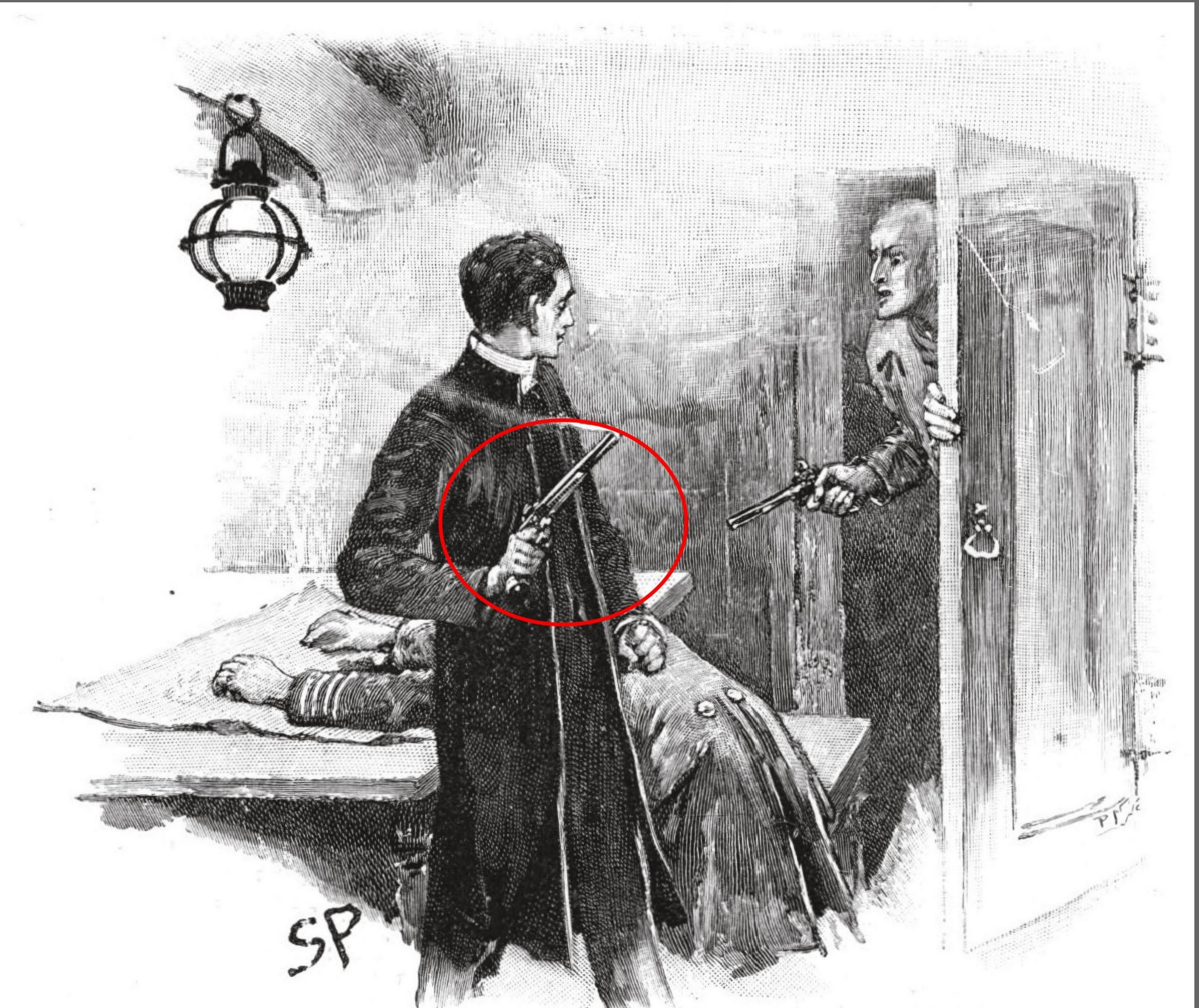
Bayesian Decision Theory

- In some models, we do have the chance to improve our understanding by making additional observations.
  - Popular examples include stock market and medical diagnosis.
  - However, there is a **cost of waiting**.
    - In the case of stock market this is equivalent to losing the chance to but a stock when it is cheaper or sell at its highest price.
    - In the case of medical diagnosis, the sickness can progress while you are waiting for additional lab results.
  - We could model this as a series of decision making windows, each one more accurate than the previous.
    - Eventually we will be always accurate because everything will be in the past. However, each time we move to the next window, we will also be losing some opportunities.

Bayesian Decision Theory

- How to model cost of waiting?
- The sequential windows of opportunity are different than each other, in the sense that there is additional information.
- At any step we are given x as the observed variables, and at the next step we are given (through further observation) a new variable z.
  - Initial window : $EU(\alpha_i|x) = \Sigma\ U_{ik}P(S_k|x)$
  - Next window : $EU(\alpha_i|x,z) = \Sigma\ U_{ik}P(S_k|x,z)$
- How about the utility?
  - If $EU(\alpha_i|x) < EU(\alpha_i|x,z)$ then our **expected utility for this particular choice** has improved.
  - If $EU(x) < EU(x,z)$ then our **overall expected utility** has improved.
  - In such cases **z is useful information, and has value**.

Bayesian Decision Theory



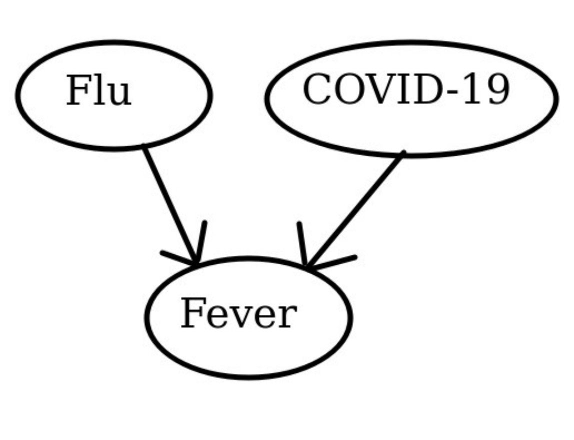
“THE CHAPLAIN STOOD WITH A SMOKING PISTOL IN HIS HAND.”

Bayesian Decision Theory

- **Bayesian Networks** (also called Belief Networks or Probabilistic Networks) are visual models for representing interactions between variables.
  - Such a network is composed of nodes and arcs (edges).
  - Each node represents a random variable (X).
  - Each node has a corresponding value P(x) which is the probability.
  - If there is an arc (directed edge) from node X to node Y, then this indicates that **X has a direct influence on Y**.
    - This also means that **Y should not have a direct influence on X**.
    - Extending on this concept, Bayesian Networks are required to be **Directed Acyclical Graphs** (DAGs) which means there can be no cycles.
      - Being a DAG is very significant from the computer science point of view.
    - The nodes and arcs are the **structure** and the probabilities are the **parameters**.

Bayesian Decision Theory

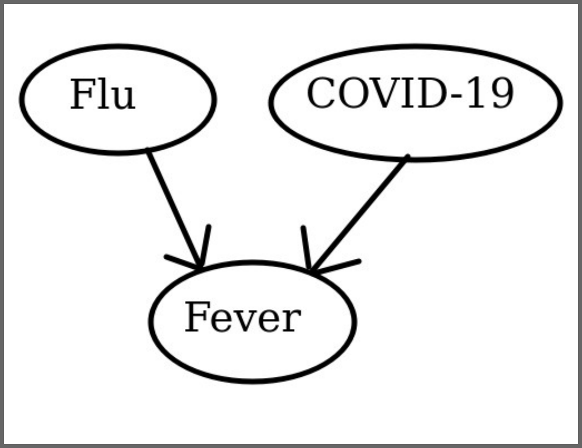
- A simple Bayesian network
  - Observed variables
    - COVID-19 has been observed in %3 of the population.
    - %40 of people without COVID-19 have fever.
      - We use flu to explain this, but we don’t have data on flu incidence rate at this moment.
    - %70 of people with COVID-19 have fever.
    - We assume no other sickness causes fever. (Important why?)
    - We assume no one gets both flu and COVID-19. (Important why?)



- P(C-19)=0.03
- P(Fever | ~C-19) = 0.40
- P(Fever | ~Flu, C-19) = 0.70
- P(Fever | ~Flu, ~ C-19) = 0

Bayesian Decision Theory

- A simple Bayesian network
  - IN the DAG, both Flu and COVID-19 have a direct influence on Fever.
  - P(C-19|Fever)
    - = P(Fever|C-19) P(C-19) / P (Fever)
    - = 0.70 x 0.03 / P (Fever)
    - = 0.021 / [P(Fever|C-19)P(C-19) + P(Fever|~C-19)P(~C-19)
    - = 0.021 / [0.70x0.03+0.40x0.97]
    - **=0.05 = %5.**
  - Knowing about fever increased COVID-19 probability from %3 to %5.
  - But not much because Flu is also explaining Fever.
  - If we get incidence rate for Flu, we can have a better model.
    - Better model does not necessarily mean higher probability for COVID-19.

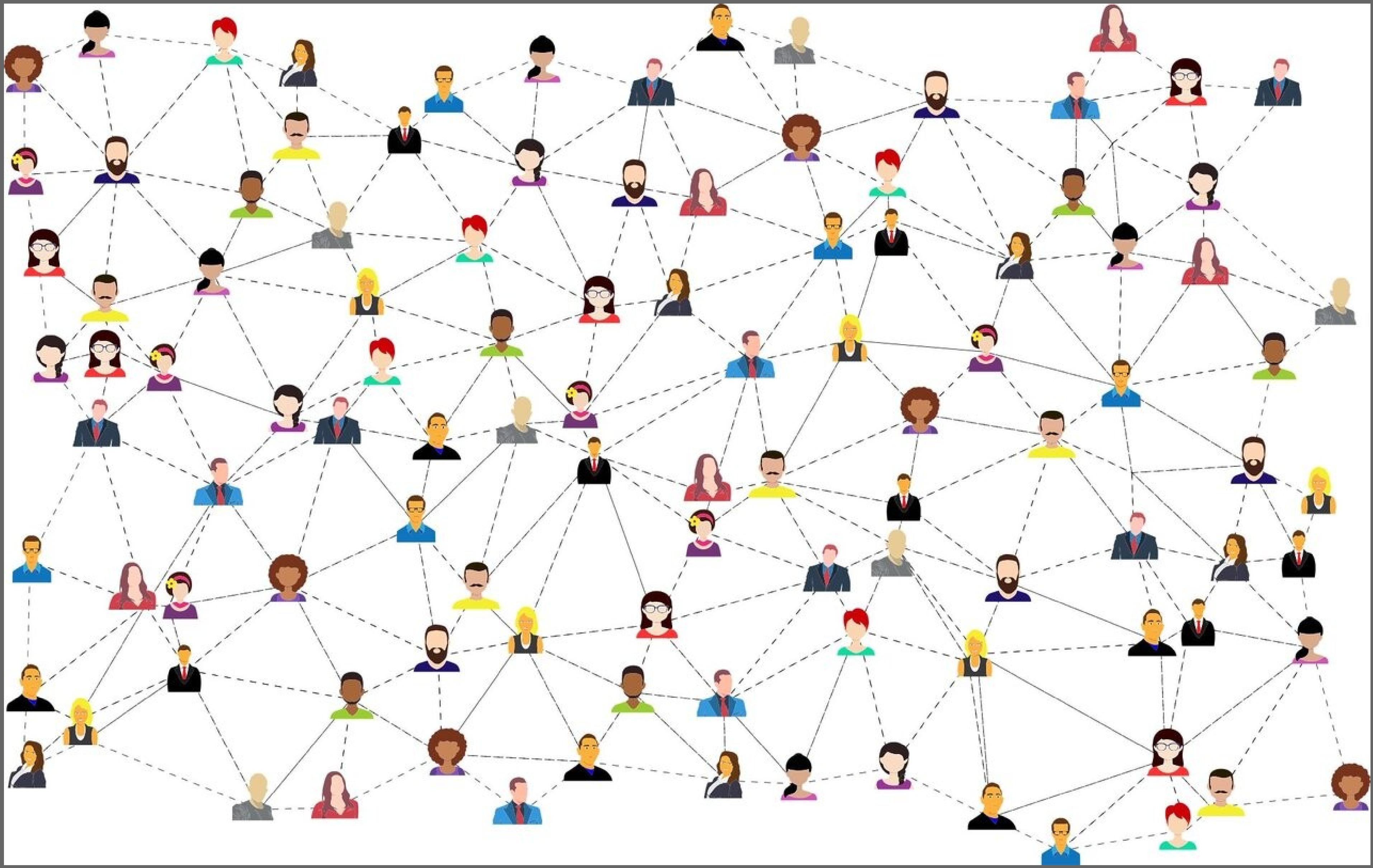


- P(C-19)=0.03
- P(Fever | Flu, ~C-19) = 0.40
- P(Fever | ~Flu, C-19) = 0.70
- P(Fever | ~Flu, ~ C-19) = 0

Bayesian Decision Theory

- In the example, we assume Flu and COVID-19 are **independent**.
  - Do we have actual scientific proof about that?
  - How about some mysterious and hypothetical factor X (genetic, nutritional, lifesytle, etc.) that has influence on both sicknesses?
  - Can we calculate the probability of having such a factor?
- Is fever the only symptom these two sicknesses have in common?
- Are there exclusive symptoms?
- The DAG would be much more complex, and the conditional probabilities could be extended further with consecutive application of Bayes' rule.
  - $P(X_1,\ldots,X_D) = \Pi\, P(X_i \mid \text{parents } (X_i))$
- Of course calculation of all these conditional probabilities becomes hard. A systematic approach using the **belief propagation algorithm** (1988) is much better. This algorithm assume that the DAG is in the form of a **tree**.
  - Many DAGs are already in the form of trees, but not all DAGs are so.
  - **Trees are special DAGs** where all nodes except one (root) has only one parent, and the root has no parents.
  - Being able to represent a DAG as a tree is beneficial. This may **require clustering some variables into one**.

Bayesian Decision Theory



Bayesian Decision Theory

- Some notes:
  - Arcs in Bayesian networks **do not necessarily imply causality**.
  - The most basic approach we discussed for classification here is called a **Naive Bayes Classifier**. It simply assumes **independent inputs**.
  - Not all probabilities are necessarily known prior to network construction. Estimating these unknown probabilities is a valuable task, but it is not easy.
- Because we cannot assume causality does not mean we cannot **estimate a confidence level for it**.
  - An association rule is in the form of **if X, then Y**.
  - The **confidence** of such rule is P (Y|X)=P(X,Y)/P(X) which is a conditional probability. It shows the strength of the rule.
  - And the **support** of such rule is P(X,Y) which is a joint probability. It shows the statistical significance of the rule.
  - The **Apriori Algorithm** (Agrawal, etal. 1996) makes multiple passes over the DAG to discover association rules with high confidence and high support.
    - When we discover these rules, we can easily infer the probability of some event based on previously observed events in the association rule.
    - Example: If you have items x, y in your basket, what is the probability of adding item z?
  - This approach is in general called **Association Rule Mining**, and is a very popular task.

Questions?

CONTACT:
bora.gungoren@atilim.edu.tr