IS590DT: Data Mining Applications

WEEK 11: NB, BN, SP and OCR

Spring 2019

Yingjun Guan
yingjun2@illinois.edu
http://ischool.illinois.edu/people/yingjun-guan

ILLINOIS | School of Information Sciences
The iSchool at Illinois

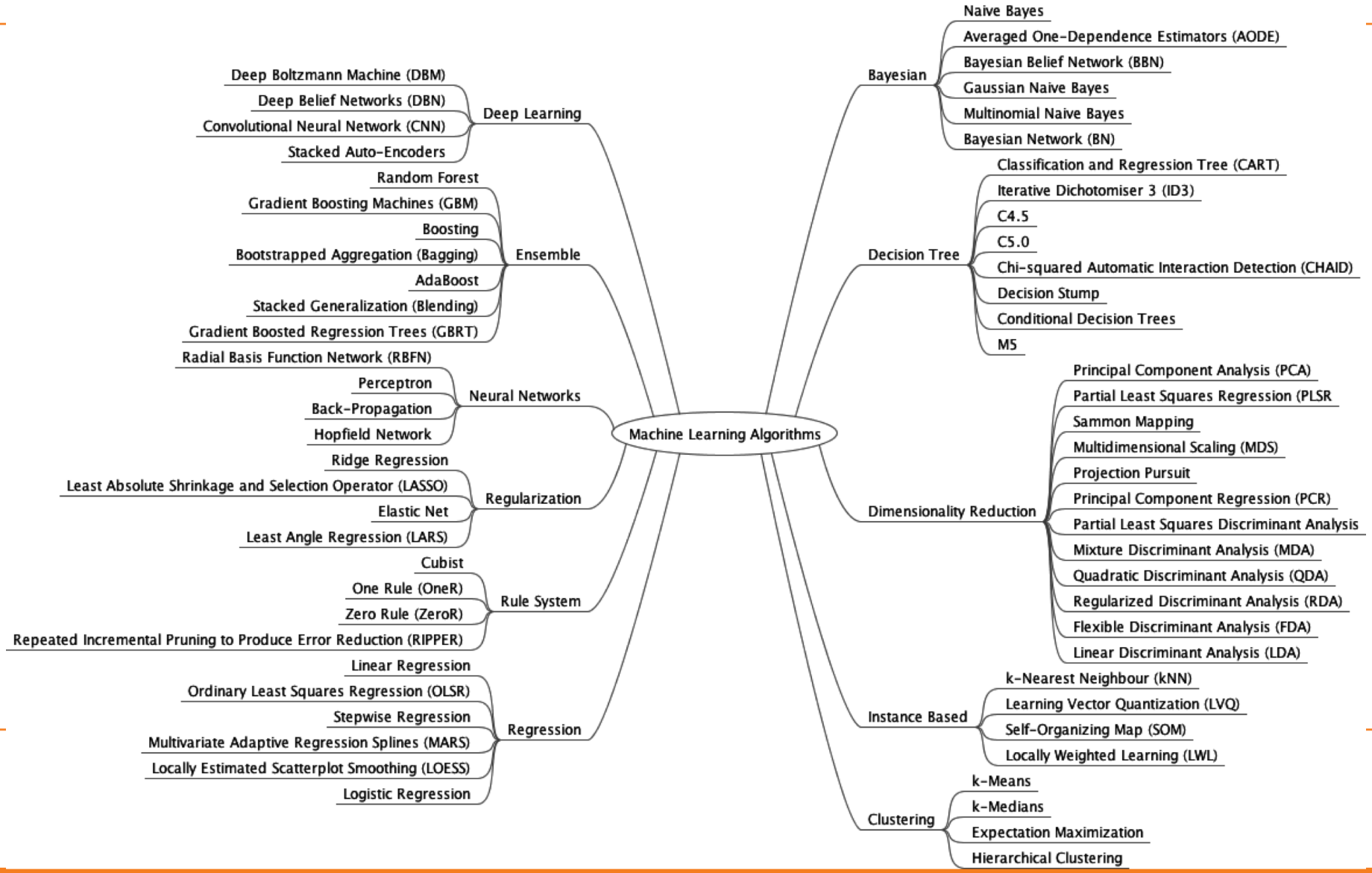
Contents today

- **Naïve Bayes**
- **Bayesian Network**
- **OCR application {Reproducibility}**
- **Structural Prediction**
- **Summary**

Let's start with how classification works…

- The classification of classification algorithms (How indeed do we make the final classification?)
- Let's check the following machine learning algorithm summarizing figure first.

Which one is classification?



Bayesian Theorem…

- Conditional probability.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Probability of event A given B has occured

Probability of event A occured and event B occured

Probability of event B

- EG: P(healthy | positive result)

Truth	Test result		Total
	Positive	Negative	
Renal disease	44	23	67
Healthy	10	60	70
Total	54	83	137

Bayesian Theorem…

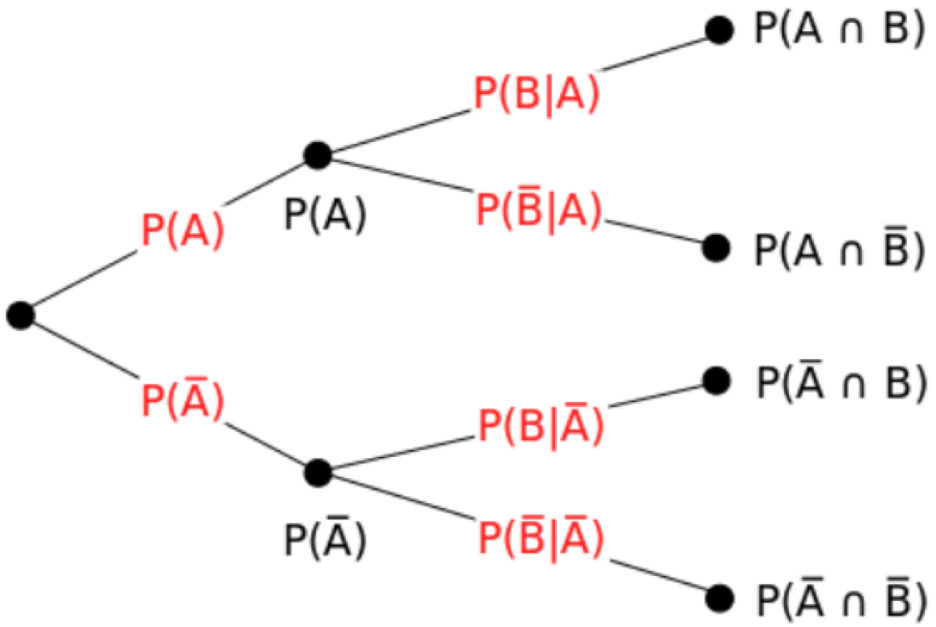
- X sometimes buys an ice cream after DT classes, depending on the the assignment Yingjun leaves.
- If it is an easy assignment (50% to happen), 90% likelihood to buy.
- If it is a hard assignment (50% to happen), 30% likelihood to buy.
- After today's class, X buys an ice cream, what do you think of today's assignment? Easy? Hard?

Formula and Tree structure

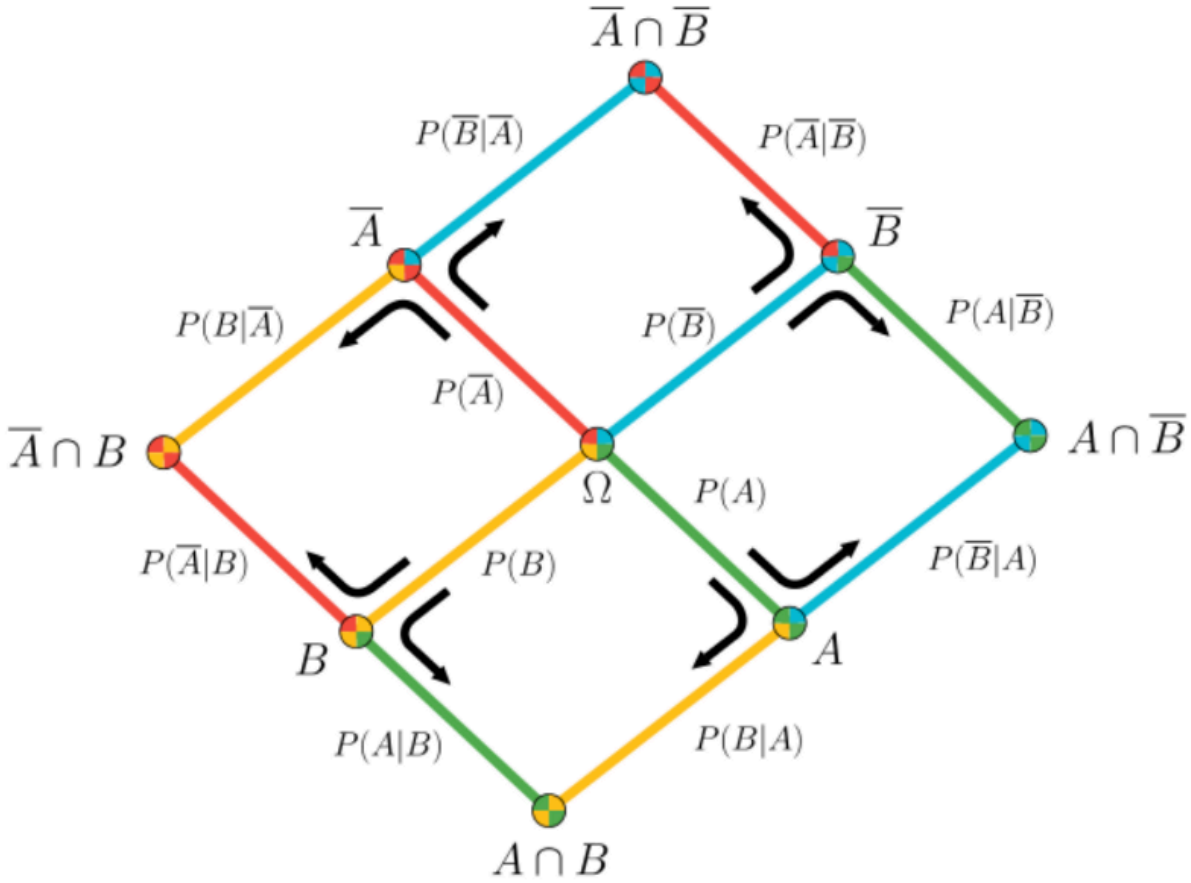
- Formula

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

- Tree Structure

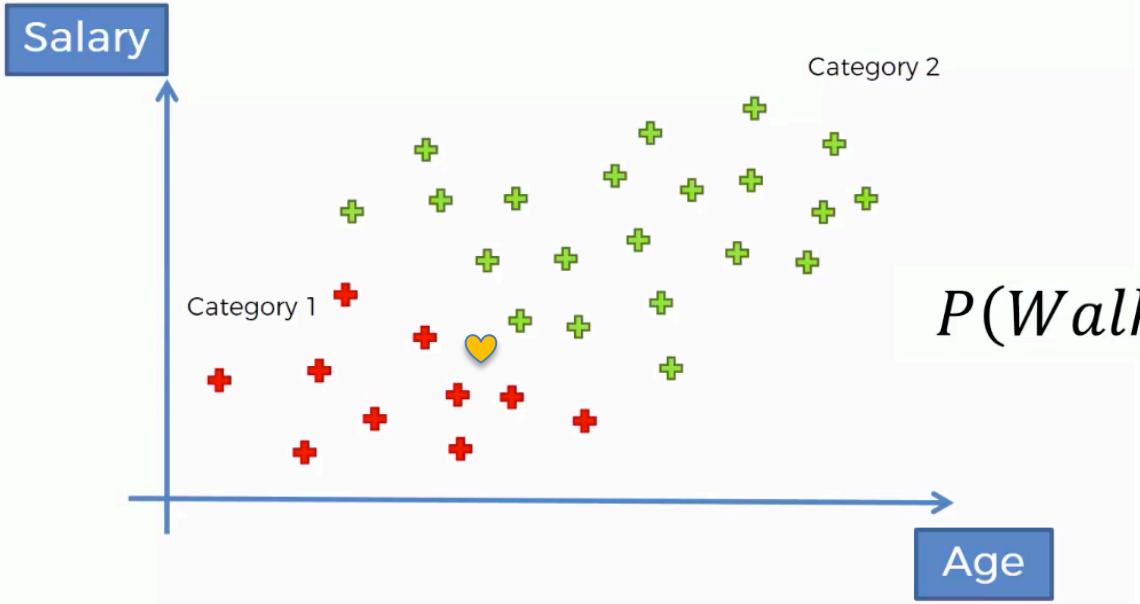


Another Visualizations of Bayes' theorem



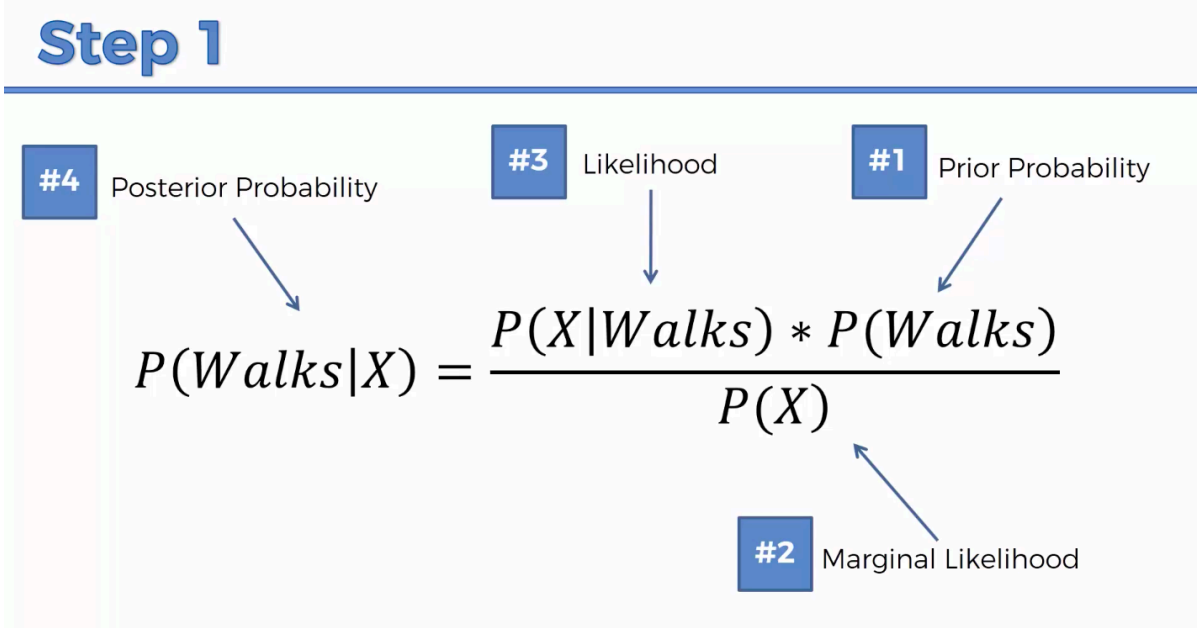
Example on naïve bayes algorithm

- What happens if we add a new observation, how do we classify it? 「walk? Drive?」

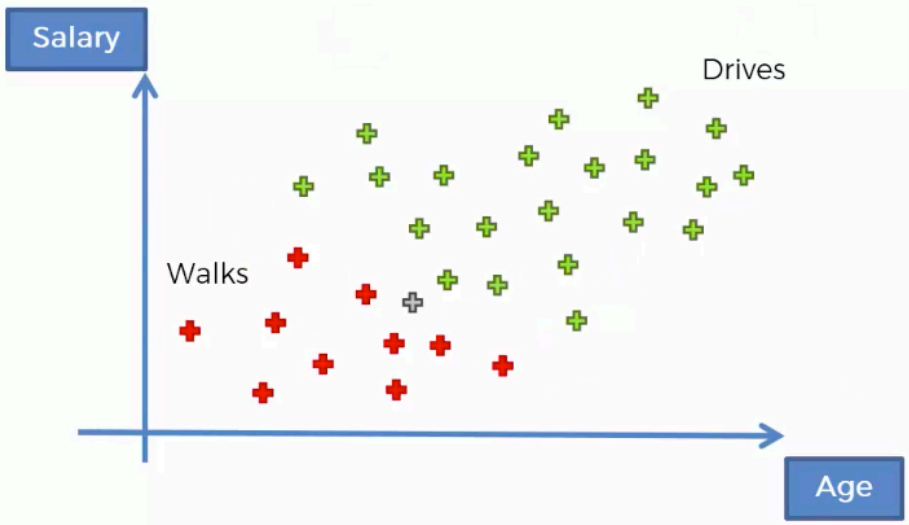


$$P(Walks|X)\ v.s.\ P(Drives|X)$$

Let's calculate P(Walks|X) first.



P(Walks|X)

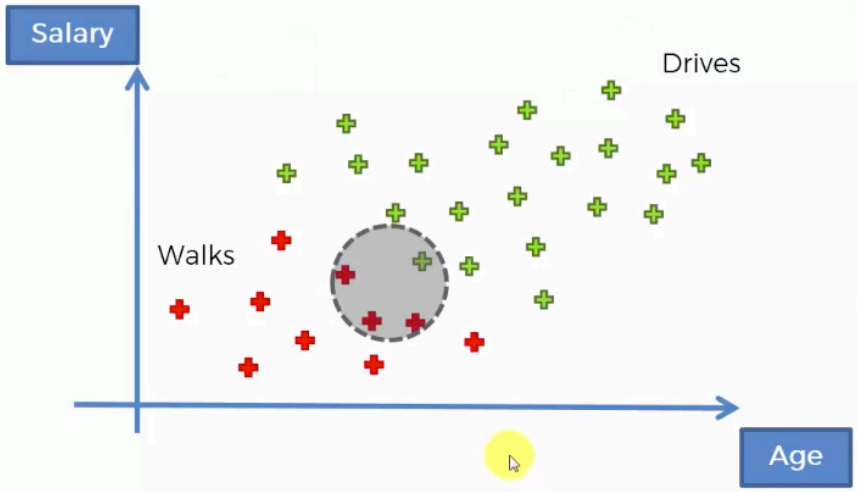


#1. P(Walks)

$$P(Walks) = \frac{Number\ of\ Walkers}{Total\ Observations}$$

$$P(Walks) = \frac{10}{30}$$

P(Walks|X)

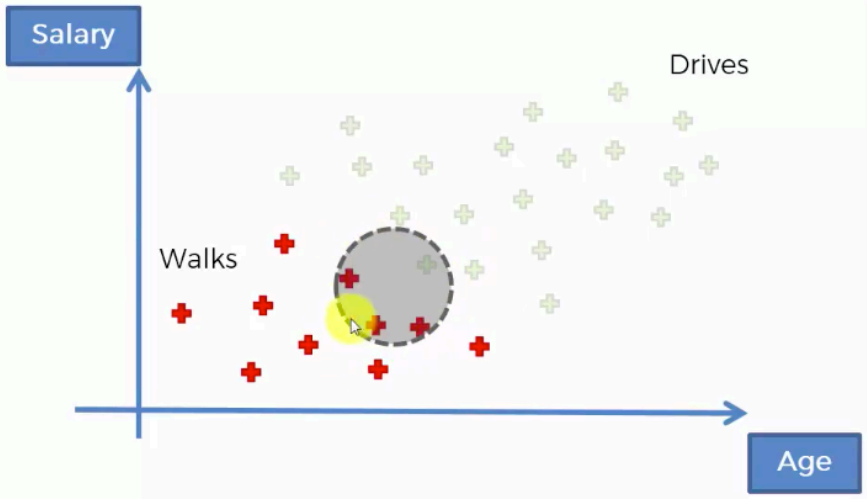


#2. P(X)

$$P(X) = \frac{Number\ of\ Similar\ Observations}{Total\ Observations}$$

$$P(X) = \frac{4}{30}$$

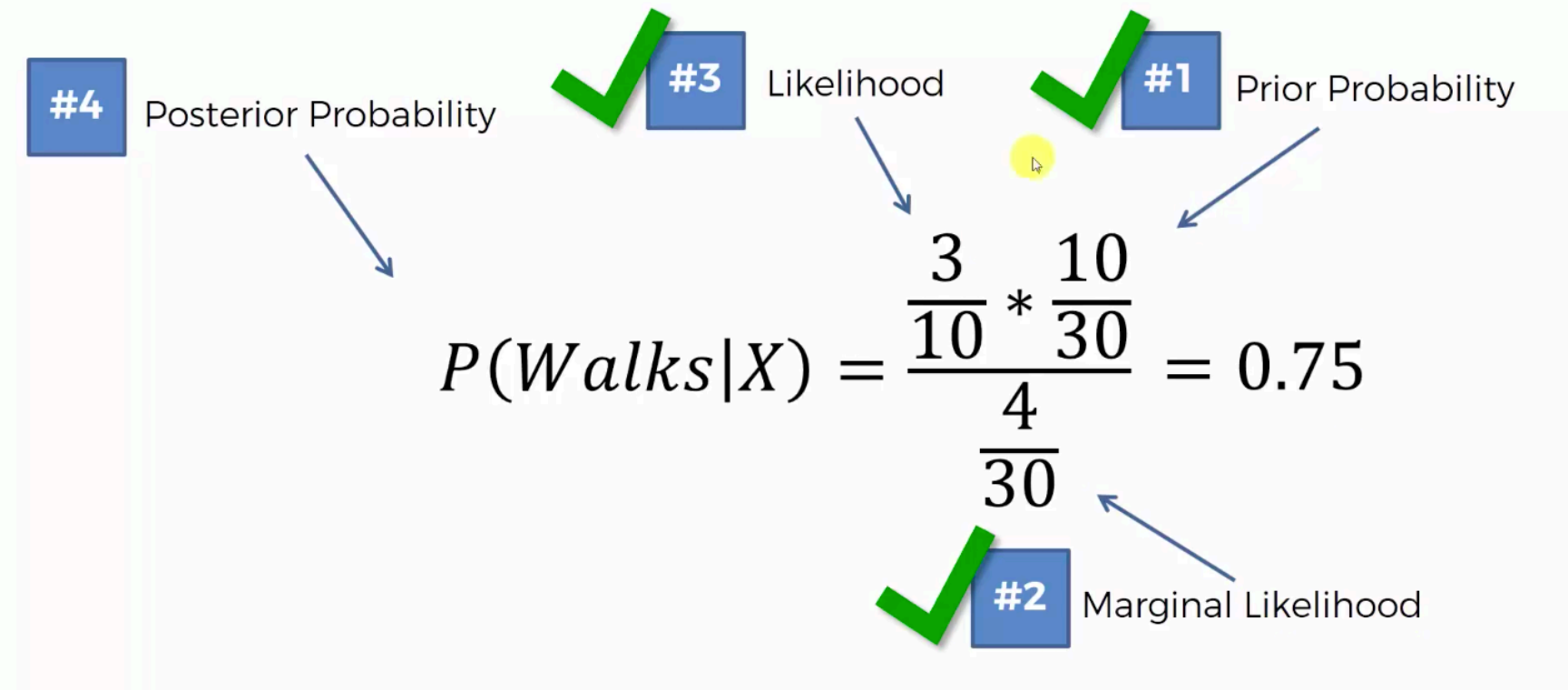
P(Walks|X)



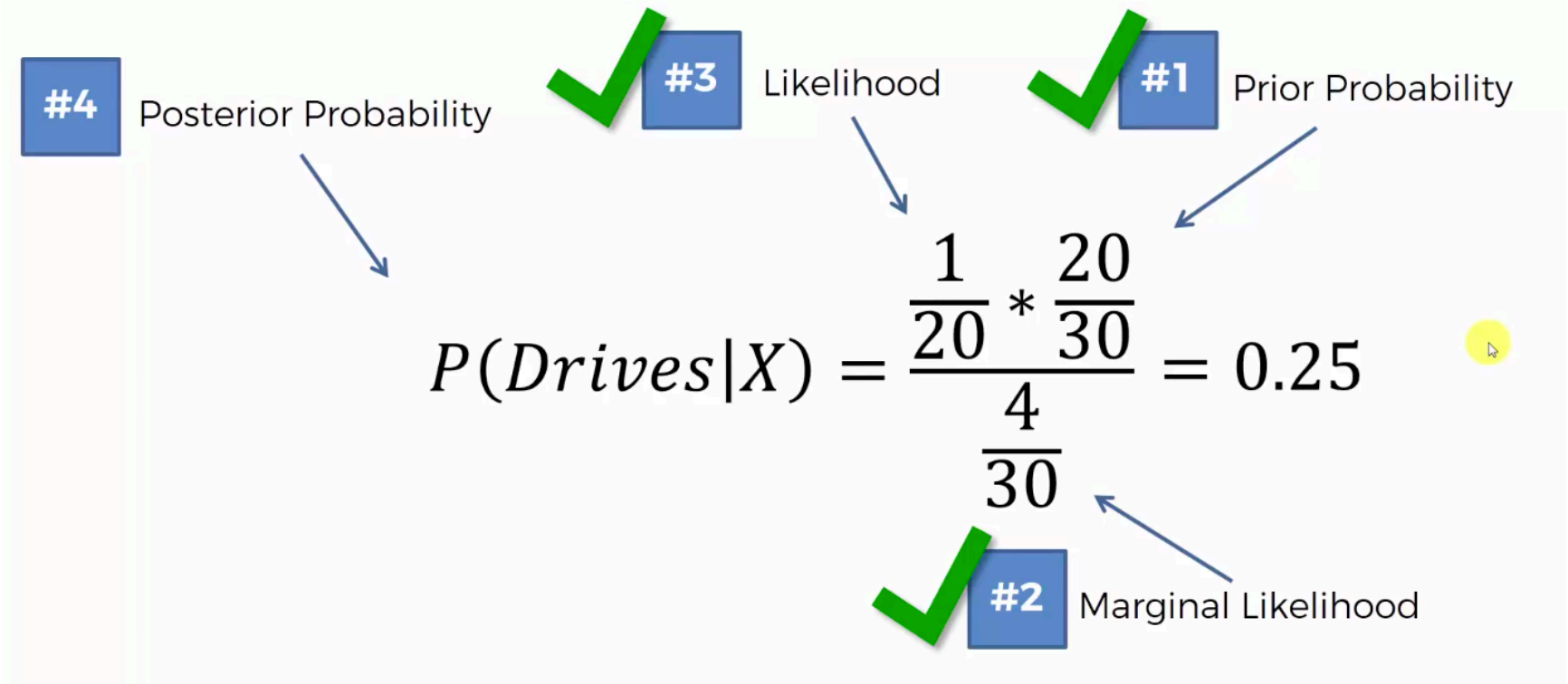
#3. P(X|Walks)

$$P(X|Walks) = \frac{Number\ of\ Similar\ Observations\ Among\ those\ who\ Walk}{Total\ number\ of\ Walkers}$$

P(Walks|X)


#4
Posterior Probability
#3
Likelihood
#1
Prior Probability
$P(Walks|X) = \frac{\frac{3}{10} * \frac{10}{30}}{\frac{4}{30}} = 0.75$
#2
Marginal Likelihood

P(Drives|X)


#4
Posterior Probability
#3
Likelihood
#1
Prior Probability
$P(Drives|X) = \frac{\frac{1}{20} * \frac{20}{30}}{\frac{4}{30}} = 0.25$
#2
Marginal Likelihood

To simplify…

$$P(Walks|X)\ v.s.\ P(Drives|X)$$

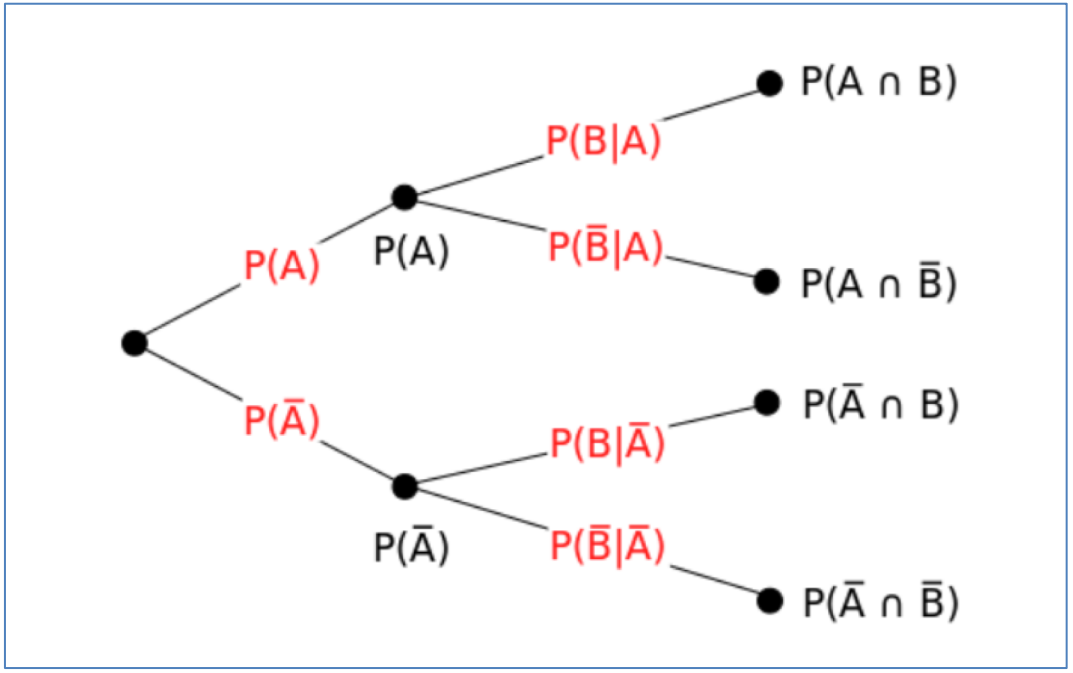
$$\frac{P(X|Walks) * P(Walks)}{\cancel{P(X)}} \quad v.s. \quad \frac{P(X|Drives) * P(Drives)}{\cancel{P(X)}}$$

When the # of classes are >2 …

- P(A1|X) = P (x|A1) * P(A1) / ?
- P(A1|X) = P (x|A1) * P(A1) / P(X)
- P(A1|X) = P (x|A1) * P(A1)
- P(A2|X) = P (x|A2) * P(A2)
- …

Why is it called Naïve Bayes?

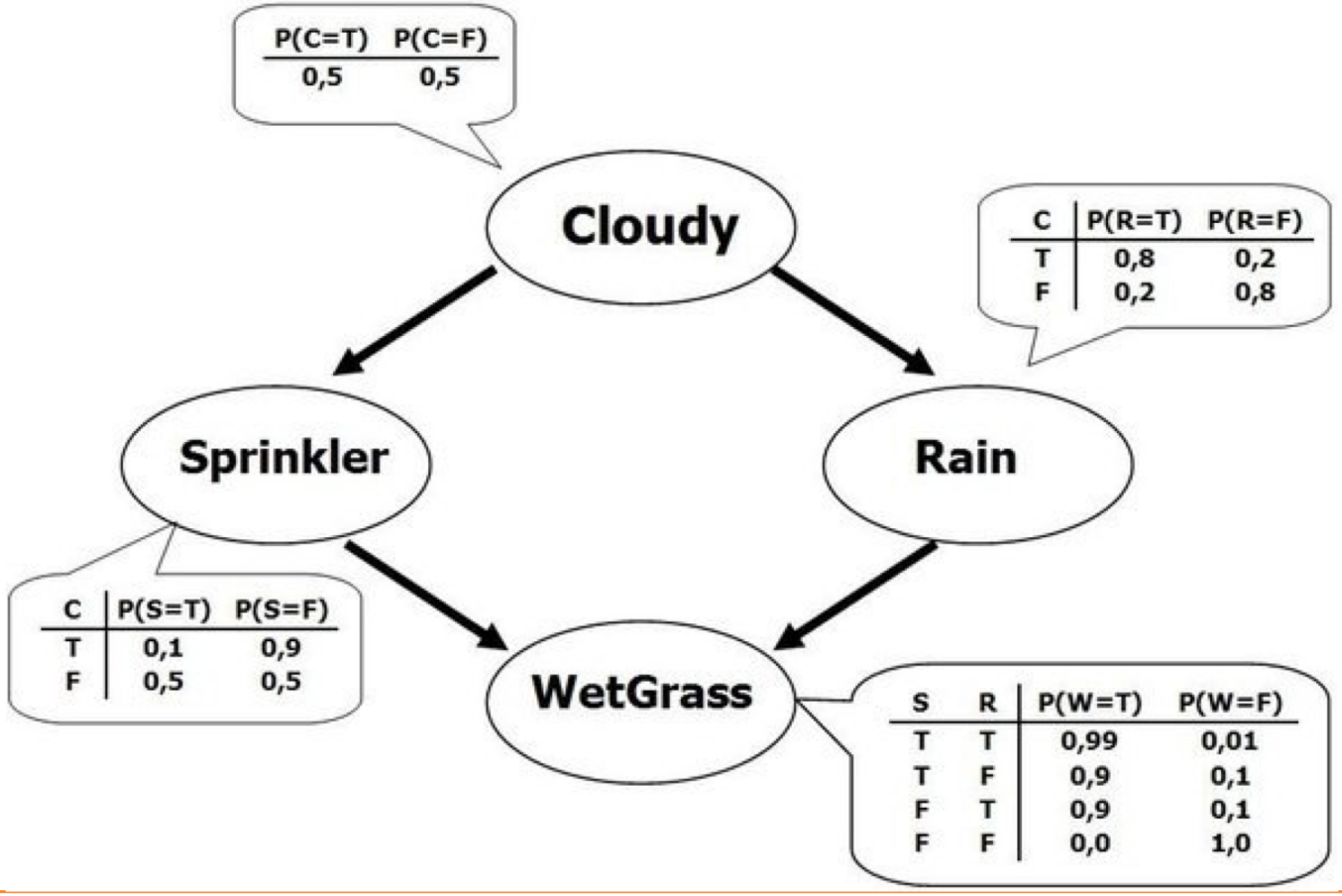
- It holds the assumption that *the attributes are conditionally independent given the class.*



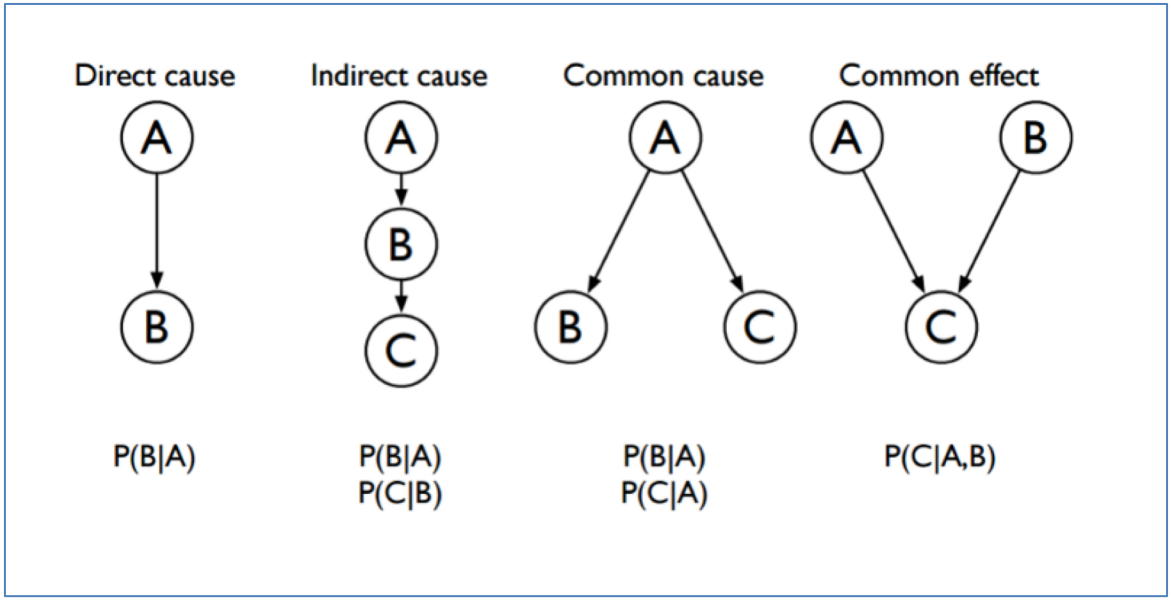
NB to BN…

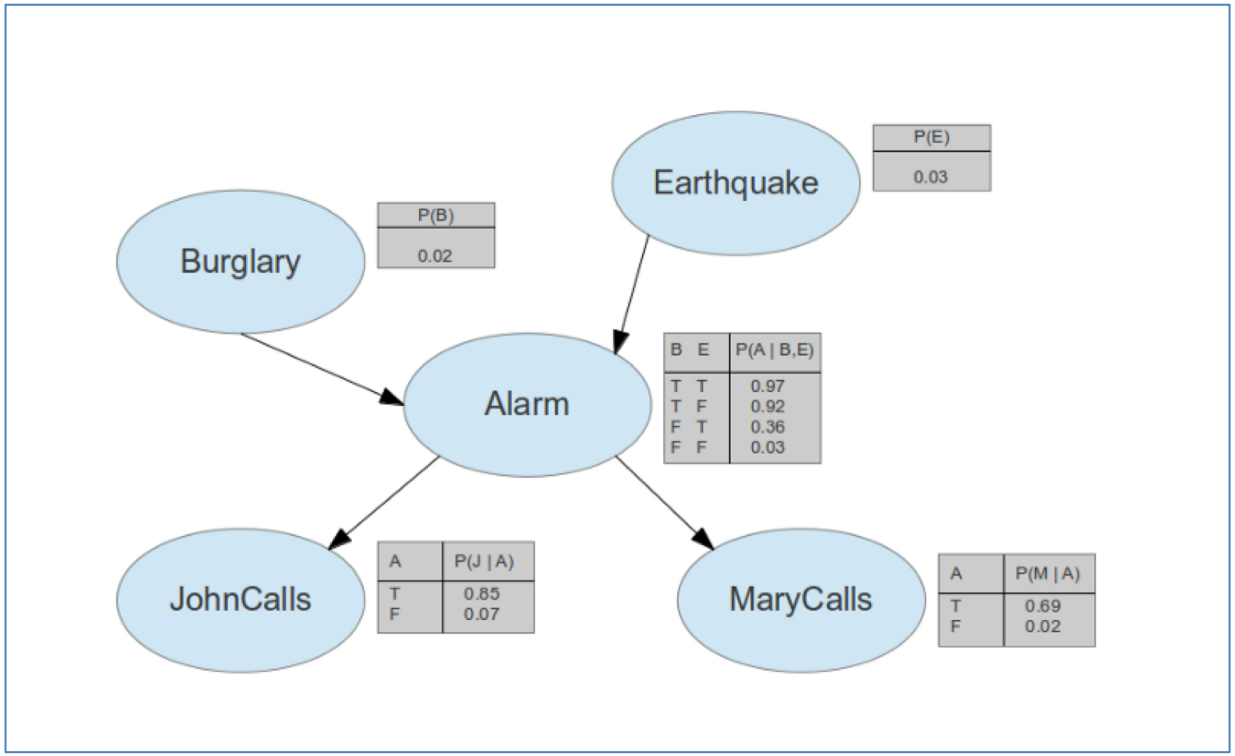
- Bayesian network is designed to eliminate the naïve assumption.

Another example



- Conditional dependency relations (arcs) from *node A* to another *node B* represent that *node B* is a child of *node A* or put it differently the *node A* is a parent of *node B*.





$$p(X_1,\cdots,X_n)=\prod_{i=1}^{n} p(X_i \mid Parents(X_i))$$

- A Bayesian network is a **directed acyclic graph** in which each edge corresponds to a conditional dependency, and each node corresponds to a unique random variable.
- DAG is used in Markov chain, Deep learning, Structural prediction, etc.

OCR case

- How to recognize OCR letters???
- How do you know it is an A?
- What info do you need to recognize it is an A?

Figure 1. Examples of the character images generated by "warping" parameters.

- Let's read a paper and see how they deal with the problem.

Letter recognition using Holland-style adaptive classifiers. PW Frey, DJ Slate - Machine Learning, 1991 – Springer.

Reading time

Letter recognition using Holland-style adaptive classifiers.
PW Frey, DJ Slate - Machine Learning, 1991 – Springer.

Read paper: try to identify

- characterization of images (attributes)
- methods used (classifiers)
- performance measures.

- Guess: what might be the most difficult letter to recognize?
- What might be the most difficult pair?
- Check your results with weka.

Let's play with the data.

Let's go to UCI website.

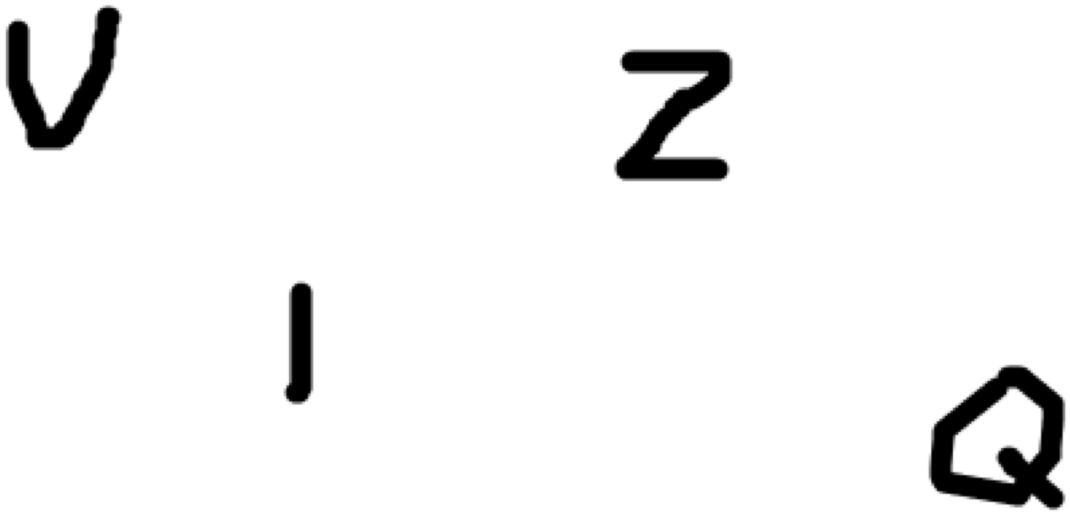
- Intro to UCI archive website.
- https://archive.ics.uci.edu/ml/datasets/Letter+Recognition
- Also abel.lis.illinois.edu/data for downloading arff file.

Reproducibility

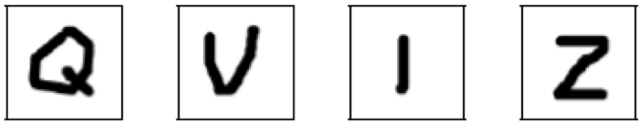
- Why is it hard to reproduce other research work?
- Replicate vs reproduce.
- Why is it important to make reproducible research outcome?
- What are the good habits to make good reproducibility?

Structural Prediction

- why is structure/are correlations useful?

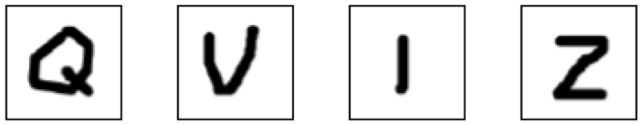


why is structure/are correlations useful?



Q V I Z

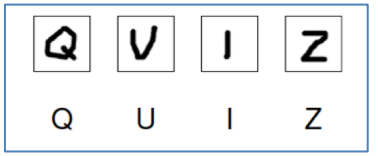
Correlations not taken into account



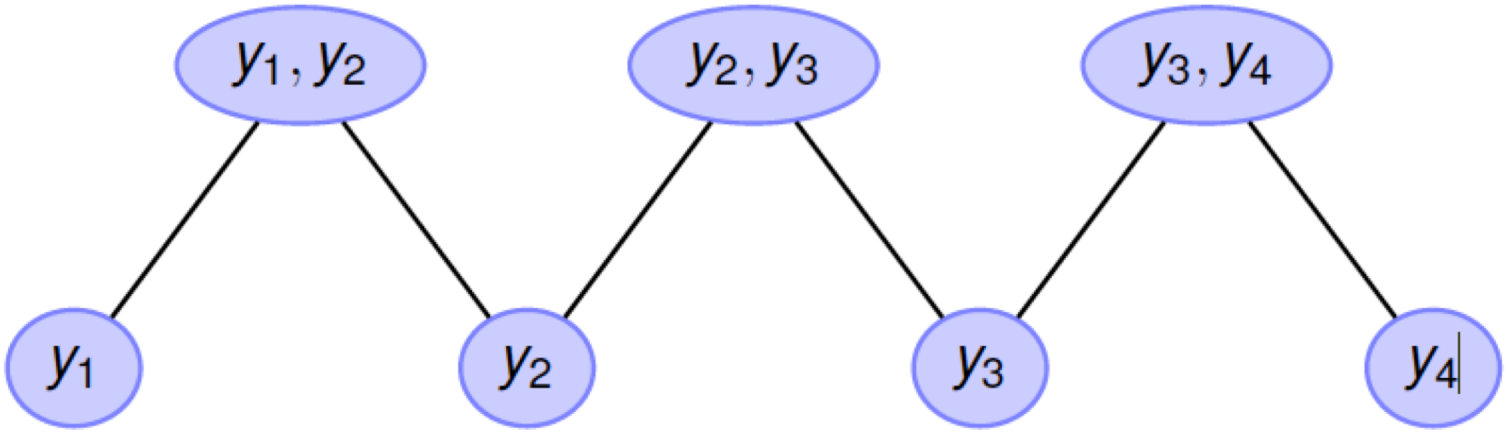
Q U I Z

Correlations taken into account

- Decomposition and visualization



$$F(\boldsymbol{w}, x, y_1, \ldots, y_4) = f_1(\boldsymbol{w}, x, y_1) + f_2(\boldsymbol{w}, x, y_2) + f_3(\boldsymbol{w}, x, y_3) + f_4(\boldsymbol{w}, x, y_4) + f_{1,2}(\boldsymbol{w}, x, y_1, y_2) + f_{2,3}(\boldsymbol{w}, x, y_2, y_3) + f_{3,4}(\boldsymbol{w}, x, y_3, y_4)$$



Structural Prediction

- Why is this helpful?
- Formulate it as prediction of all four letter words (multiclass prediction):

$$y \in \mathcal{Y} = \{1, \ldots, 26^4\}$$

- Problem: Really large output space A.
- Using correlations (eg: dictionary), the output space is much smaller.

Why and when do we need SP?

- Another example: denoising.

Predictions from neighboring pixel are useful.

- SP is very powerful in image processing.
- You might have heard of CNN, RNN, deep learning, etc.
- SP is their very good foundation.

- Limitation of SP.

Let's go through the assignment

- Another OCR problem.

Mid-term feedback

- Complete the feedback in moodle for the course.
- Q & A time

- That’s all for today.