Bayesian Classifiers

(with a recap of probabilistic representation)

Applied Machine Learning (EDAN95)
Lecture 10
2019-12-04
Elin A. Topp

Material based on Lecture Slides on Probabilistic Representation and Bayesian Learning, EDAF70, Spring 2018,
Lecture 11, EDAN95 Fall 2018
Goodfellow et al, "Deep Learning", and Russel/Norvig, "AI - A Modern Approach"

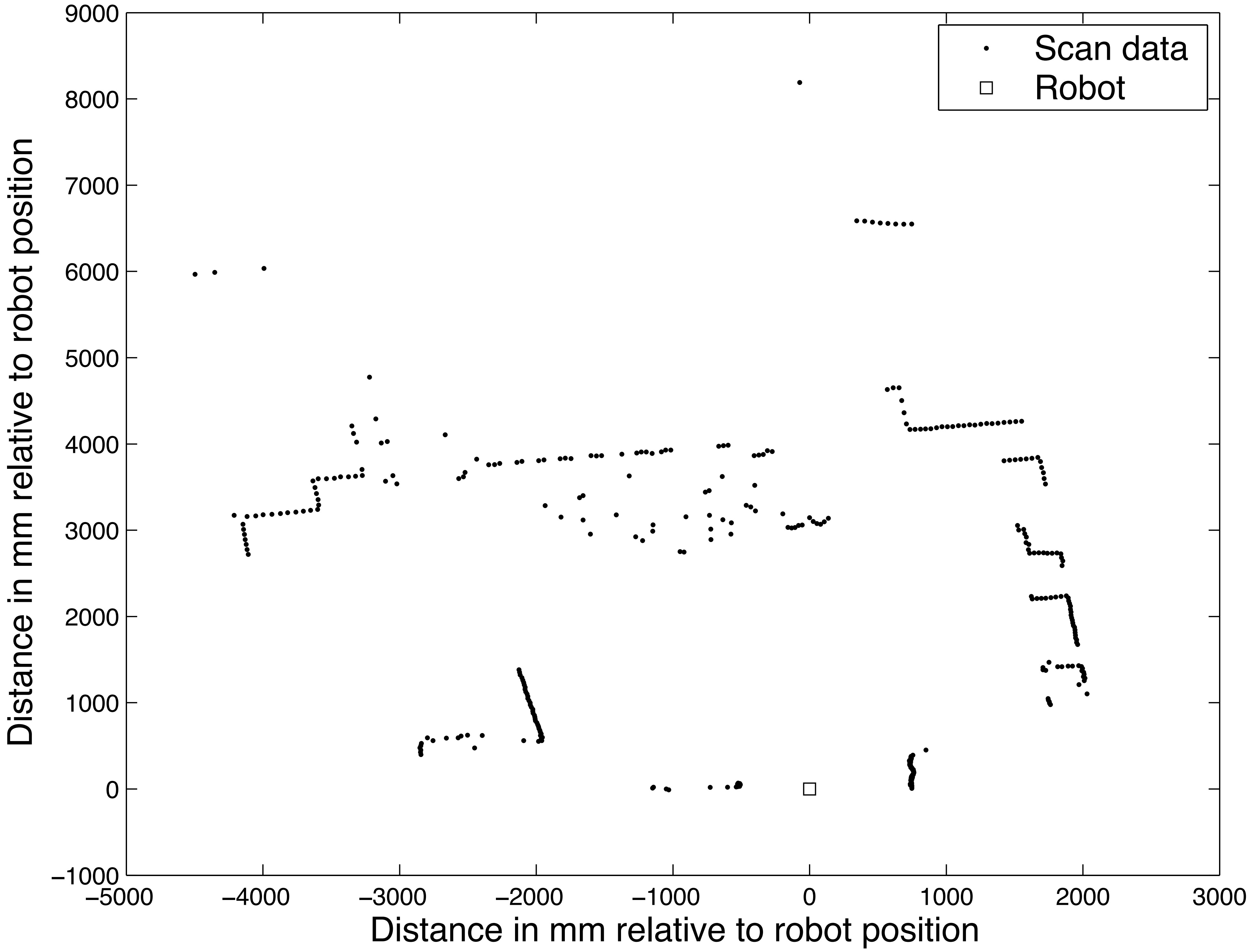
Today’s agenda

- Recap conditional / posterior probabilities, Bayes’ rule, independence / conditional independence
- Brief introduction to Bayesian Networks (BN)
- The Naive Bayesian Classifier as special case of a BN
- Learning a Bayesian Classifier
- Gaussian Mixture Models

Today’s agenda

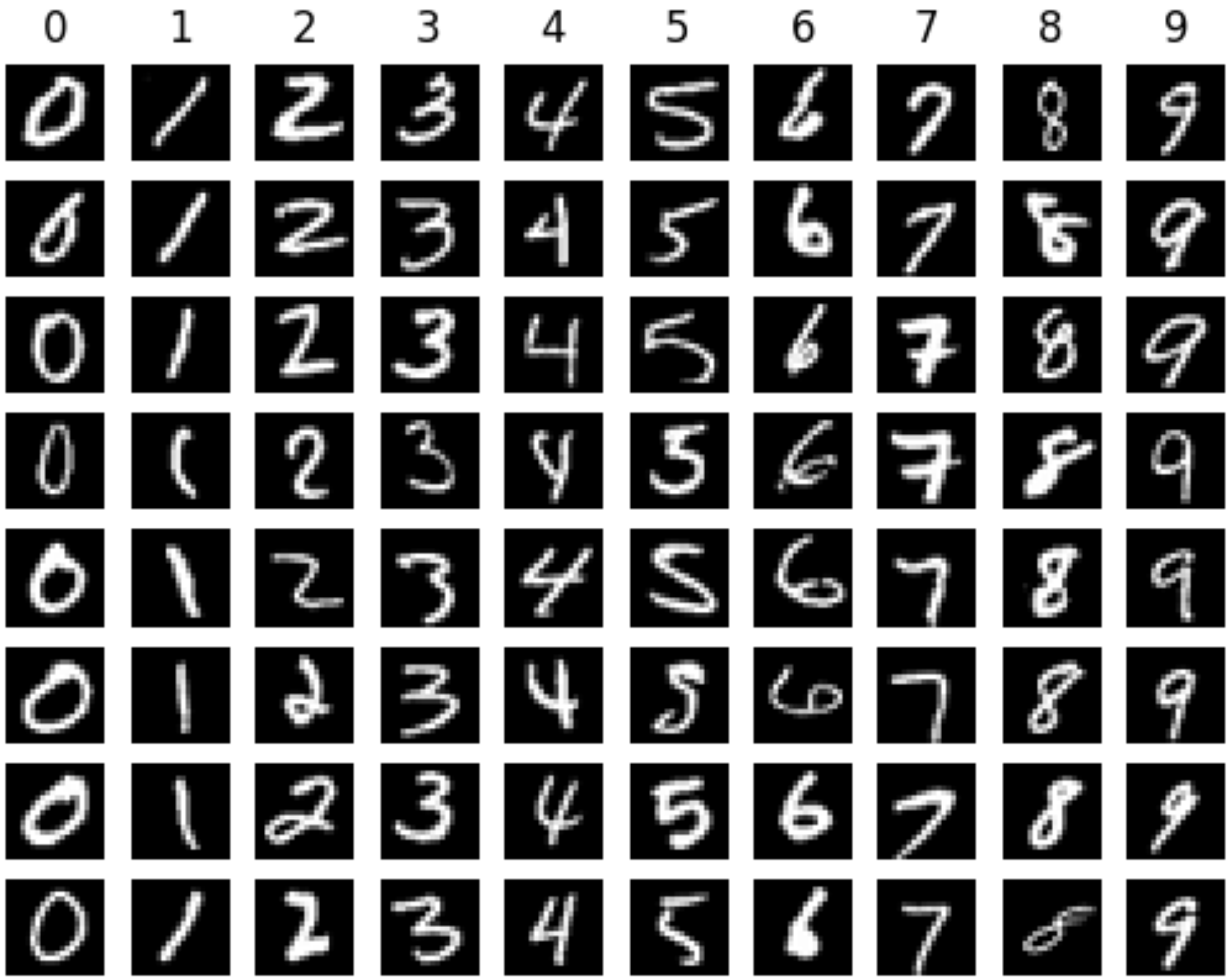
- Recap conditional / posterior probabilities, Bayes’ rule, independence / conditional independence
- Brief introduction to Bayesian Networks (BN)
- The Naive Bayesian Classifier as special case of a BN
- Learning a Bayesian Classifier
- Gaussian Mixture Models

A robot’s view of the world...



Which “leg-like” data point patterns were caused by a person’s leg, which by furniture?

Or back to MNIST



Which combination of pixel values are most often seen for each of the numbers?
Which number is it that best explains the pixel values of a specific sample?

Bayesian learning

We want to classify / categorize / label new observations based on experience

More general: We want to *predict* and *explain* based on (limited) experience, to find categories / labels for observations or even the model for “how things work” (transition models, sensor models) *given a series of (explained) observations*.

First needed: Recap on conditional probabilities!

Bayesian Probability

Probabilistic assertions summarise effects of

laziness: failure to enumerate exceptions, qualifications, etc.

ignorance: lack of relevant facts, initial conditions, etc.

Subjective or Bayesian probability:

Probabilities relate propositions to one's state of knowledge (*A = "the observed pattern in the data was caused by a person"*)

e.g., *P(A) = 0.2*

e.g., *P(A | there is a ton of "leggy" furniture in the respective room) = 0.1*

Not claims of a "probabilistic tendency" in the current situation, but maybe learned from past experience of similar situations.

Probabilities of propositions change with new evidence:

e.g., *P(A | ton of furniture, dataset obtained at 7:30 by a bot) = 0.05*

Some notations

We express propositions as random variables taking on certain values directly

We look then for example at

$P(X = x_i), i = 1,\dots\ n$, for all n values x_i of the Variable X

Thus: $P(X = x_1) = P(X = x_2) = 1/2$

with e.g., x_1 = “dice roll outcome is odd number” and x_2 = “dice roll outcome is even number”

For the *distribution* over the possible values of X we get then:

$\mathbb{P}(X) = < P(X = x_1), P(X = x_2), \dots, P(X = x_n) >$

and we use vector notation $\mathbf{P}(X)$ to indicate that we iterate over a subset of the values for X in a computation of a joint distribution, e.g.

$\mathbb{P}(X,Y) \quad = \quad \mathbb{P}(X\,|Y)\ \mathbf{P}(Y)$ describes a set of equations, expressing the joint probability distribution of X and Y as conditional probability distribution of X in dependency of the possible (or specifically given) values of Y

Prior probability

Prior or *unconditional probabilities* of propositions

e.g., $P(\mathit{Person} = \mathit{true}) = 0.2$ and

$P(\mathit{Weather} = \mathit{sunny}) = 0.72$ (e.g., known from statistics)

correspond to belief *prior to the arrival of any (new) evidence*

Probability distribution gives values for all possible assignments (normalised):

$\mathbb{P}(\mathit{Weather}) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$

Joint probability distribution for a set of (*independent*) random variables gives the probability of every atomic event on those random variables (i.e., every sample point):

$\mathbb{P}$(Weather, Person) = $\mathbb{P}$(Weather) X $\mathbb{P}$(Person), i.e., a 4 x 2 matrix of values:

Weather Person	sunny	rain	cloudy	snow
true	0.144	0.02	0.016	0.02
false	0.576	0.08	0.064	0.08

Posterior probability

Most often, there is *some* information, i.e., *evidence*, that one can base their belief on:

e.g., *P(person) = 0.2* (prior, no evidence for anything), but

P(person | leg-size) = 0.6

or

P(number = 0) = 0.1 (in a uniformly distributed subset of MNIST-data), but

P(number = 0 | pixel[36] = black) = 0.6 (rough, educated guess based on the digits data)

corresponds to belief *after the arrival of some evidence* (also: *posterior* or *conditional probability*).

OBS: NOT "if leg-size, then 60% chance of person"

THINK *"given that leg-size is all I know" instead!*

Evidence remains valid after more evidence arrives, but it might become less useful

Evidence may be completely useless, i.e., irrelevant.

P(person | leg-size, sunny) = P(person | leg-size)

Domain knowledge lets us do this kind of inference.

Posterior probability (2)

Definition of conditional / posterior probability:

$$P(a \mid b) = \frac{P(a \wedge b)}{P(b)} \quad \text{if } P(b) \neq 0$$

or as *Product rule* (for a and b being true, we need b true and then a true, given b):

$$P(a \wedge b) \quad = \quad P(a \mid b)\, P(b) \quad = \quad P(b \mid a)\, P(a)$$

and in general (independency cannot be assumed) for whole distributions (e.g.):

$$\mathbb{P}(\mathit{Weather}, \mathit{Person}) \quad = \quad \mathbb{P}(\mathit{Weather} \mid \mathit{Person})\, \boldsymbol{P}(\mathit{Person})$$

(a *4x2* set of equations, governed by the chosen (given) value for Person from the array over possible values, hence ***P***)

Chain rule (successive application of product rule):

$$\mathbb{P}(X_1, ..., X_n) = \mathbb{P}(X_1, ..., X_{n-1})\, \mathbb{P}(X_n \mid X_1, ..., X_{n-1})$$

$$= \mathbb{P}(X_1, ..., X_{n-2})\, \mathbb{P}(X_{n-1} \mid X_1, ..., X_{n-2})\, \mathbb{P}(X_n \mid X_1, ..., X_{n-1})$$

$$= ... = \prod_{i=1}^{n} \mathbb{P}(X_i \mid X_1, ..., X_{i-1})$$

Bayes' Rule

Recap *product rule:* $P(a \wedge b) = P(a \mid b)\, P(b) = P(b \mid a)\, P(a)$

$$\Rightarrow \text{ Bayes' Rule } P(a \mid b) = \frac{P(b \mid a)\, P(a)}{P(b)}$$

or in distribution form (vector notation to express, that for the distribution, we normally look at all possible outcomes for Y that govern P(X)):

$$\mathbf{P}(Y \mid X) = \frac{\mathbf{P}(X \mid Y)\, \mathbf{P}(Y)}{\mathbf{P}(X)} = \alpha\, \mathbf{P}(X \mid Y)\, \mathbf{P}(Y)$$

Useful for assessing *diagnostic* probability from *causal* probability

$$P(\mathit{cause} \mid \mathit{effect}) = \frac{P(\mathit{effect} \mid \mathit{cause})\, P(\mathit{cause})}{P(\mathit{effect})}$$

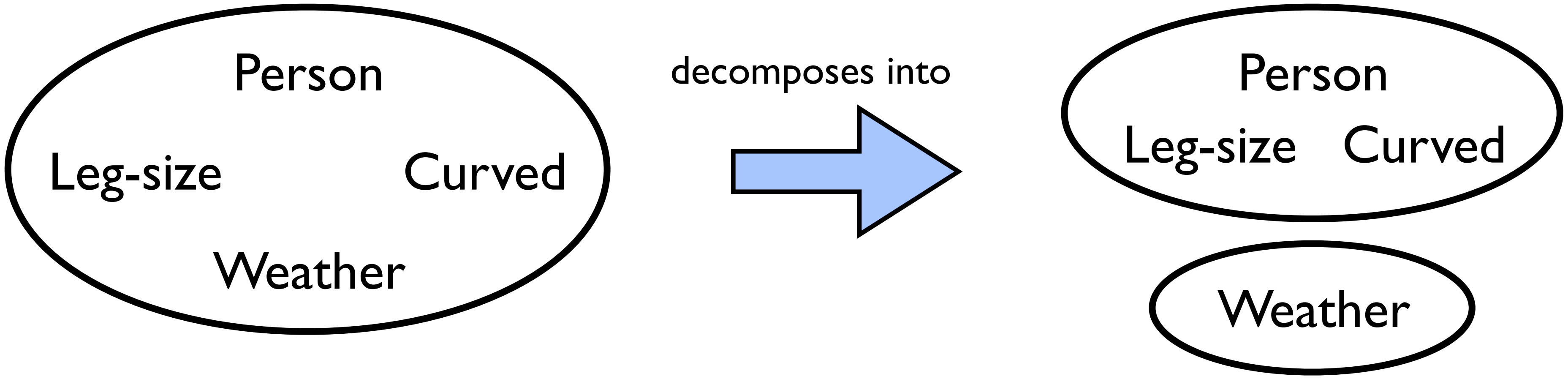
E.g., with *M* "meningitis", *S* "stiff neck":

$$P(m \mid s) = \frac{P(s \mid m)\, P(m)}{P(s)} = \frac{0.7 * 0.00002}{0.01} = 0.0014$$ (not too bad, really!)

Independence

A and B are *independent*, i.e., $A \perp B$ iff

$$P(A \mid B) = P(A) \quad \text{or} \quad P(B \mid A) = P(B) \quad \text{or} \quad P(A, B) = P(A)\, P(B)$$



Person

Leg-size

Curved

Weather

decomposes into

Person

Leg-size

Curved

Weather

$$\mathbb{P}(\mathit{Leg\text{-}size}, \mathit{Curved}, \mathit{Person}, \mathit{Weather}) = \mathbb{P}(\mathit{Leg\text{-}size}, \mathit{Curved}, \mathit{Person})\ \mathbb{P}(\mathit{Weather})$$

32 entries reduced to 8 + 4 (Weather is not Boolean!).
This absolute (*unconditional*) independence is powerful but rare!

Some fields (like robotics and computer vision, or, as used in the AIMA book, dentistry) have still a lot, maybe hundreds, of variables, none of them being independent.

What can be done to overcome this mess...?

Conditional independence

$\mathbb{P}(\textit{Leg-size}, \textit{Person}, \textit{Curved})$ has $2^3 - 1 = 7$ independent entries (must sum up to 1)

But: If there is a person, the probability for "Curved" does not depend on whether the pattern has leg-size (this dependency is now "implicit" in some sense):

$$(1)\ \mathbb{P}(\textit{Curved} \mid \textit{leg-size}, \textit{person}) = \mathbb{P}(\textit{Curved} \mid \textit{person})$$

The same holds when there is no person:

$$(2)\ \mathbb{P}(\textit{Curved} \mid \textit{leg-size}, \neg\textit{person}) = \mathbb{P}(\textit{Curved} \mid \neg\textit{person})$$

Curved is *conditionally independent* of *Leg-size* given *Person*:

$$\mathbb{P}(\textit{Curved} \mid \textit{Leg-size}, \textit{Person}) = \mathbb{P}(\textit{Curved} \mid \textit{Person})$$

Writing out the full joint distribution using chain rule:

$$\begin{aligned}
&\mathbb{P}(\textit{Leg-size}, \textit{Curved}, \textit{Person})\\
&= \mathbb{P}(\textit{Leg-size} \mid \textit{Curved}, \textit{Person})\ \mathbb{P}(\textit{Curved}, \textit{Person})\\
&= \mathbb{P}(\textit{Leg-size} \mid \textit{Curved}, \textit{Person})\ \mathbb{P}(\textit{Curved} \mid \textit{Person})\ \mathbb{P}(\textit{Person})\\
&= \mathbb{P}(\textit{Leg-size} \mid \textit{Person})\ \mathbb{P}(\textit{Curved} \mid \textit{Person})\ \mathbb{P}(\textit{Person})
\end{aligned}$$

gives thus $2 + 2 + 1 = 5$ independent entries

Today's agenda

- Recap conditional / posterior probabilities, Bayes' rule, independence / conditional independence
- Brief introduction to Bayesian Networks (BN)
- The Naive Bayesian Classifier as special case of a BN
- Learning a Bayesian Classifier
- Gaussian Mixture Models

Bayesian networks

A simple, graphical notation for *conditional independence assertions* and hence for compact specification of full joint distributions

Syntax:

a set of nodes, one per random variable

a directed, acyclic graph (link ≈ "directly influences")

a conditional distribution for each node given its parents:

$\mathbb{P}(X_i \mid Parents(X_i))$

In the simplest case, conditional distribution represented as a

conditional probability table (CPT)

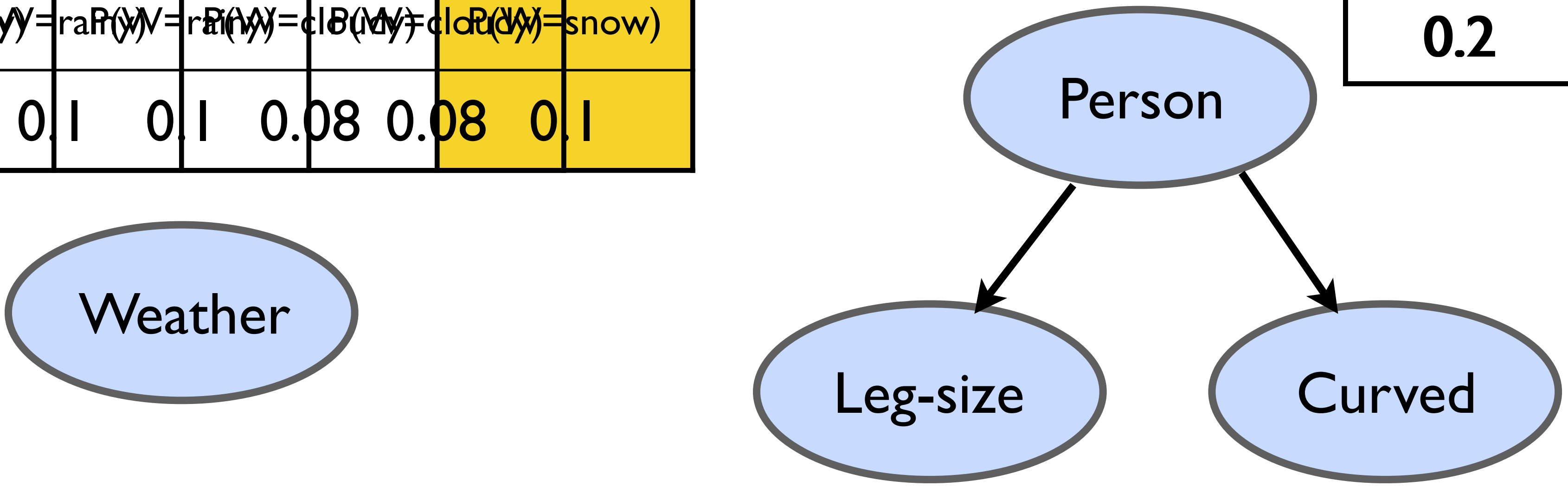
giving the distribution over X_i for each combination of parent values

Example

Topology of network encodes conditional independence assertions:

P(W=sunny)	P(W=sunny)	P(W=rainy)	P(W=rainy)	P(W=cloudy)	P(W=cloudy)	P(W=snow)
0.72	0.72	0.1	0.1	0.08	0.08	0.1

P(Per)	P(¬Per)
0.2	0.8



Per	Per	P(L|Per)	P(L|Per)	P(¬L|Per)
T	T	0.6	0.6	0.4
F	F	0.1	0.1	0.9

Per	Per	P(C|Per)	P(C|Per)	P(¬C|Per)
T	T	0.9	0.9	0.1
F	F	0.2	0.2	0.8

Weather is (unconditionally, absolutely) independent of the other variables

Leg-size and *Curved* are conditionally independent given *Person*

We can skip the dependent columns in the tables to reduce complexity!

Today's agenda

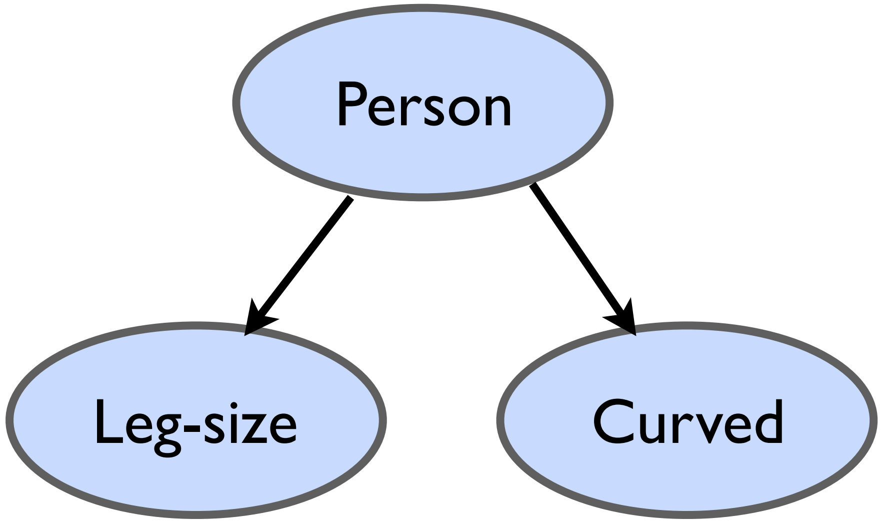
- Recap conditional / posterior probabilities, Bayes' rule, independence / conditional independence
- Brief introduction to Bayesian Networks (BN)
- The Naive Bayesian Classifier as special case of a BN
- Learning a Bayesian Classifier
- Gaussian Mixture Models

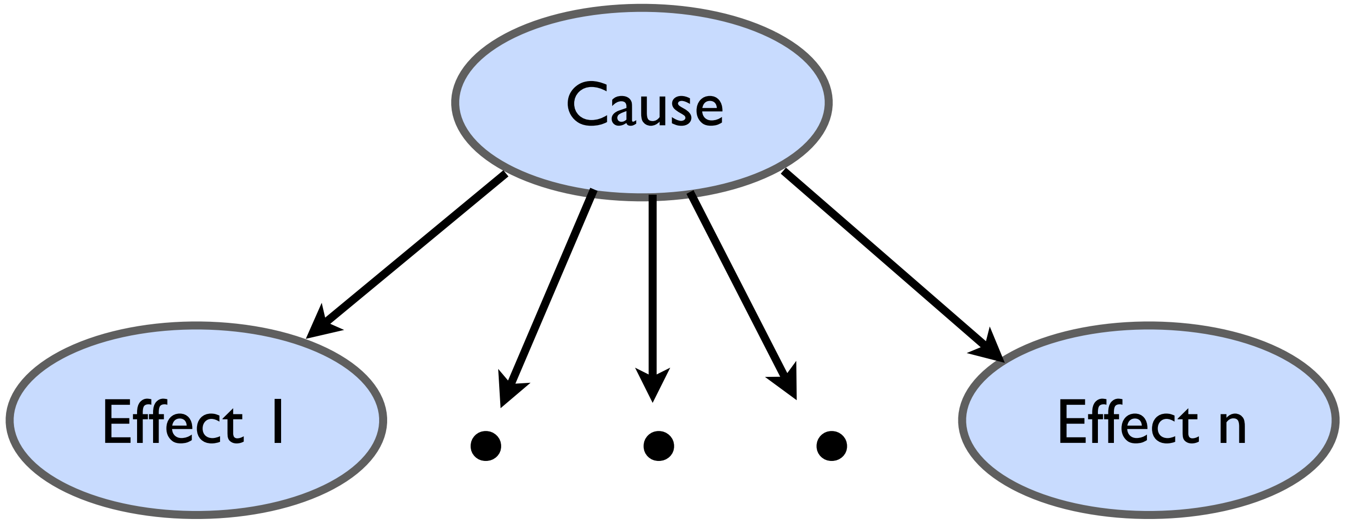
Bayes' Rule and conditional independence

$$\begin{aligned}
&\mathbb{P}(\, \mathit{Person} \mid \mathit{leg\text{-}size} \wedge \mathit{curved}) \\
&= \alpha \;\; \mathbb{P}(\, \mathit{leg\text{-}size} \wedge \mathit{curved} \mid \mathit{Person})\; \mathbb{P}(\, \mathit{Person}) \\
&= \alpha \;\; \mathbb{P}(\, \mathit{leg\text{-}size} \mid \mathit{Person})\mathbb{P}(\, \mathit{curved} \mid \mathit{Person})\; \mathbb{P}(\, \mathit{Person})
\end{aligned}$$

An example of a *naive Bayes* model:

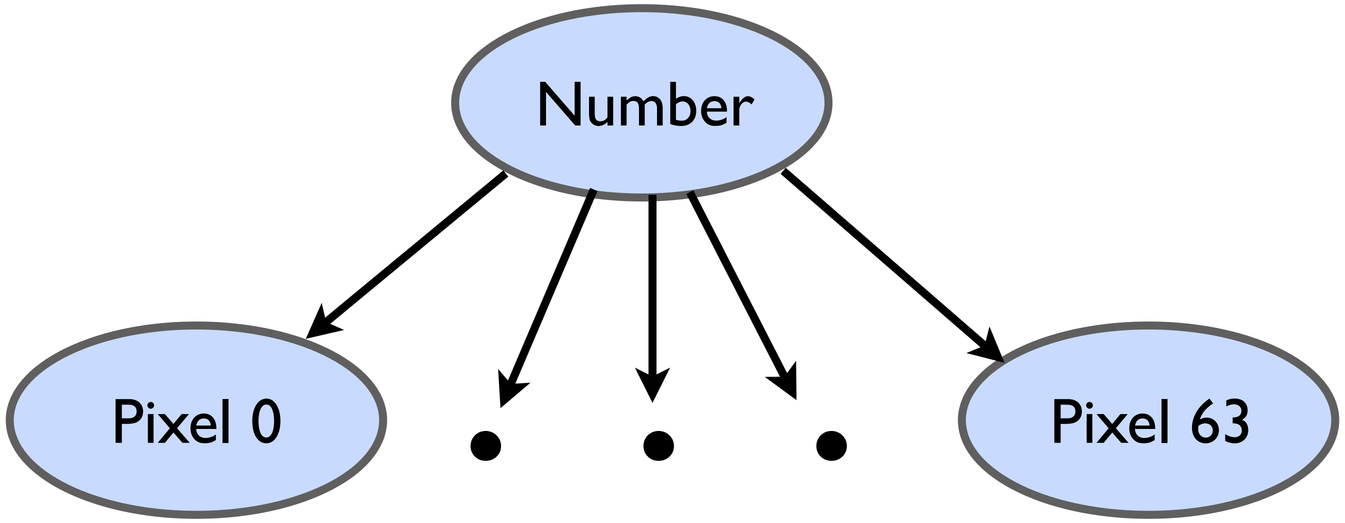
$$\mathbb{P}(\, \mathit{Cause}, \mathit{Effect}_1, \ldots, \mathit{Effect}_n) = \mathbb{P}(\, \mathit{Cause}) \prod_i \mathbb{P}(\, \mathit{Effect}_i \mid \mathit{Cause})$$





The total number of parameters is *linear* in *n*

A (super naive) NBC for the digits data



$CPT_{jk} = P(Pixel_0 = v_j | Number = k)$

$CPT_{jk} = P(Pixel_{63} = v_j | Number = k)$

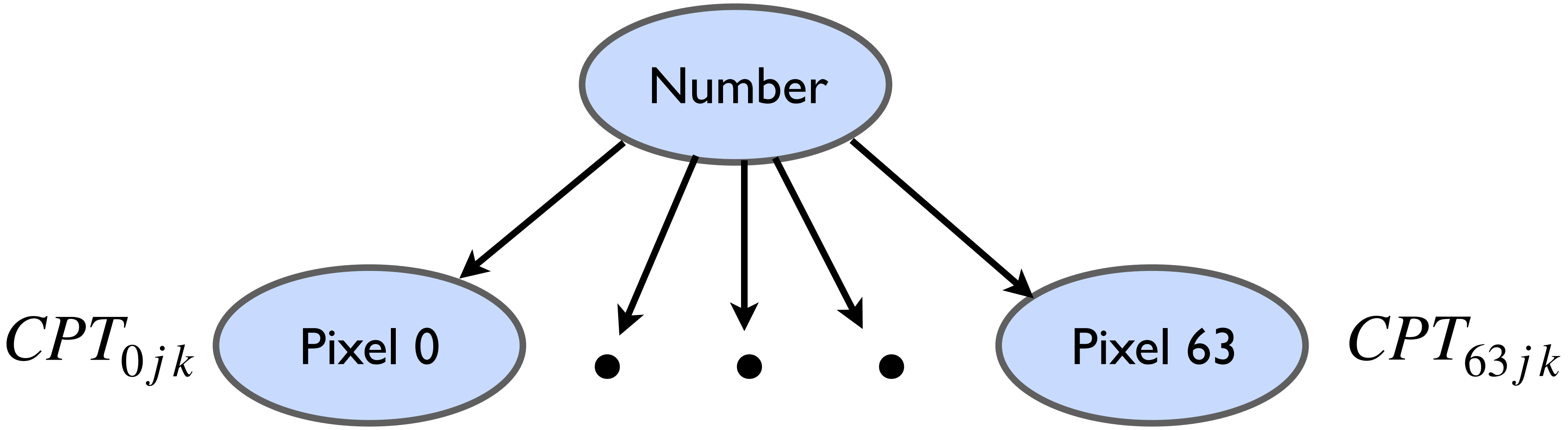
$$P(X = Number) = \arg\max_k [P(Number = k, Pixel_0, \ldots, Pixel_{n-1})]$$

$$= \arg\max_k [P(Number = k) \prod_i P(Pixel_i = X_i | Number = k)]$$

Today's agenda

- Recap conditional / posterior probabilities, Bayes' rule, independence / conditional independence
- Brief introduction to Bayesian Networks (BN)
- The Naive Bayesian Classifier as special case of a BN
- Learning a Bayesian Classifier
- Gaussian Mixture Models

Towards a (less naive) NBC



Super naive assumption was $CPT_{ijk} = P(Pixel_i = v_j \,|\, Number = k) = \dfrac{|X_{i\,v_j}|}{|X_k|}$

with $X_{iv_j} = \{examples\, X\, belonging\, to\, class\, k, where\, Pixel_i(X) = v_j\}$

Does that really work well?

- The digits data set has images with 8x8 = 64 pixels with discrete values in the range [0, …, 16]
- The more realistic MNIST_Light data has 20x20 = 400 pixels with (discrete) values in the range [0.0, …, 255.0]
- Is there a more general way to express how much an example belongs to a class?

- Assign "padded" values with the m-estimate to avoid empty slots (see the text example)
- "Blur" the probabilities by using a suitable distribution (often, not always, a Gaussian Normal Distribution can help)

Excourse: Classifying text

Our approach to representing arbitrary text is disturbingly simple: Given a text document, such as this paragraph, we define an attribute for each word position in the document and define the value of that attribute to be the English word found in that position. Thus, the current paragraph would be described by 111 attribute values, corresponding to the 111 word positions. The value of the first attribute is the word "our", the value of the second attribute is the word "approach", and so on. Notice that long text documents will require a larger number of attributes than short documents. As we shall see, this will not cause us any trouble. (*)

$$v_{NB} = \underset{v_j \in \{\text{like, dislike}\}}{argmax}\, P(v_j) \prod_i^{111} P(a_i \mid v_j) = P(v_j)\, P(a_1 = \text{“our”} \mid v_j) * \ldots. * P(a_{111} = \text{“trouble”} \mid v_j)$$

(*)[Tom M. Mitchell, "Machine Learning", p 180]

Naive Bayes Classifier for text

Given a test person who classified 1000 text samples into the categories “like” and “dislike” (i.e., the target value set *V*) and those text samples (*Examples*), the text from the previous slide is to be classified with the help of the Naive Bayes Classifier. This algorithm (from Tom M. Mitchell, “Machine Learning”, p 183) assumes (and learns) the *m-estimate* for $P(w_k | v_j)$, the term describing the probability that a randomly drawn word from a document in class v_j will be the word w_k.

```
LEARN_NAIVE_BAYES_TEXT( Examples, V)
/* learn probability terms $P( w_k | v_j)$ and the class prior probabilities $P( v_j )$ */
1. Collect all words, punctuation, and other tokens that occur in Examples
    • Vocabulary ⟵ the set of all distinct words and other tokens occurring in any text document from Examples
2. calculate the required $P( v_j )$ and $P( w_k | v_j)$ terms
    • $docs_j$ ⟵ the subset of documents from Examples for which the target value is $v_j$
    • $P( v_j ) \longleftarrow | docs_j | / | Examples |$
    • $Text_j$ ⟵ a single document created by concatenating all members of $docs_j$
    • $n$ ⟵ total number of distinct word positions in $Text_j$
    • for each word $w_k$ in Vocabulary
        • $n_k$ ⟵ number of times word $w_k$ occurs in $Text_j$
        • $P( w_k | v_j) \longleftarrow ( n_k +1) / ( n + | Vocabulary |)$          /* m-estimate */
```

```
CLASSIFY_NAIVE_BAYES_TEXT( Doc)
/* Return the estimated target value for the document Doc. $a_i$ denotes the word found in ith position within Doc.
    • positions ⟵ all word positions in Doc that contain tokens found in Vocabulary
    • Return $v_{NB}$, where
```

$$v_{NB} = \operatorname*{argmax}_{v_j \in V} \; P(v_j) \prod_{i \in positions} P(a_i | v_j)$$

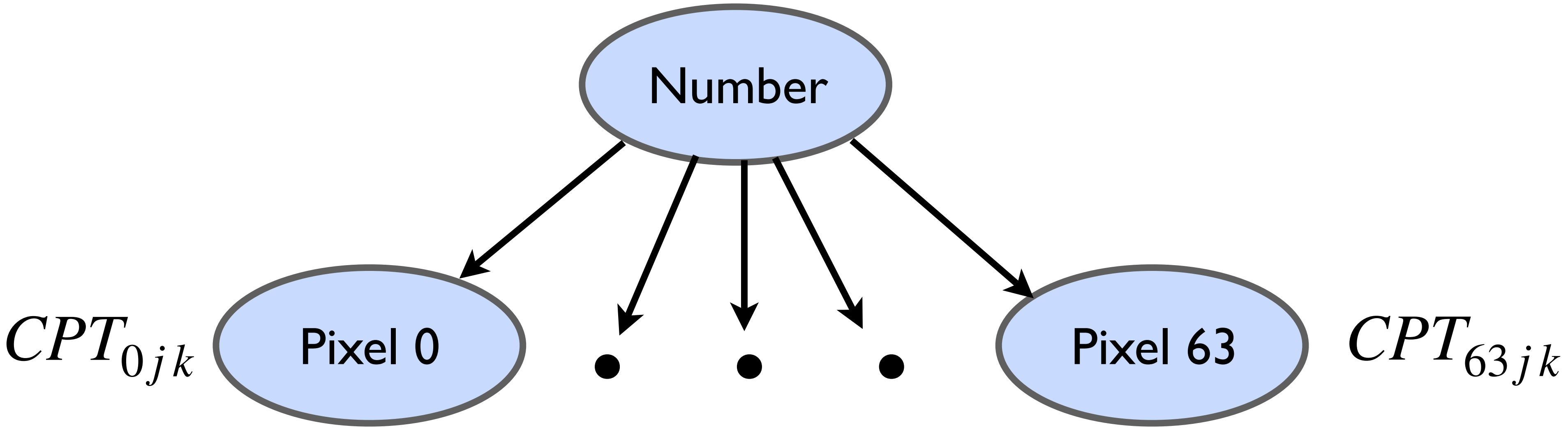
Today's agenda

- Recap conditional / posterior probabilities, Bayes' rule, independence / conditional independence
- Brief introduction to Bayesian Networks (BN)
- The Naive Bayesian Classifier as special case of a BN
- Learning a Bayesian Classifier
- Gaussian Mixture Models

Gaussian Mixture Model

- Assume that the *n* attributes in the example set form the axes of an *n*-dimensional feature space, i.e., each example is a "point" in that space.
- The examples belonging to a class will then somehow "gather" around some centre "point"
- The degree of "belonging" can be expressed as a continuous PDF - very often a Gaussian Normal distribution is suitable, which gives then a Gaussian Mixture Model (the multidimensional bell "curves" will most likely overlap, hence, there is a mixture of several distributions that explain a given data point - the sample to be classified).
- see Goodfellow (3) / Murphy (2) for other "standard" distributions

Gaussian Naive Bayesian Classifier



$$CPT_{ik} = (\mu_{ik}, \sigma_{ik}), \quad P(Pixel_i = x \,|\, Number = k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} e^{-\frac{1}{2\sigma_{ik}^2}(x-\mu_{ik})^2}$$

with $\mu_{ik} = mean(Pixel_i(X))$ and $\sigma_{ik}^2 = var(Pixel_i(X)) \quad \forall X \in \{examples\ where\ Number = k\}$

Classification is then handled for unseen sample X_{new}:

$$P(X_{new} = Number) = \arg\max_k \,[P(Number = k, Pixel_0, \ldots, Pixel_{n-1})]$$

$$= arg \max_k \,[P(Number = k) \prod_i P(Pixel_i(X_{new}) \,|\, Number = k)]$$

Outlook on lab 5

- Several implementations of statistical classifiers (NCC, NBC, and GNB)
- Several data sets
- Make sure that you can show / discuss results for any required combination of data set and classifier, do not overwrite anything as you go …

- Do not panic regarding the timing - the code for one arbitrary classifier is in the ballpark of < 55 LoC, with ~15 LoC being the core part that would have to be adapted to the respective classifier.
- No report to be delivered for this lab session - but be prepared that the content might be relevant to the next report assignment connected to lab session 6.

Today's summary

- Refreshed memory on conditional probabilities, Bayes' rule, independence, conditional independence
- [Gave a short intro / recap to Bayesian Networks (see Russel / Norvig, ch 14)]
- Introduced Naive Bayesian Classifiers
- Introduced Gaussian Mixture Models and GNBs (very briefly)

- Reading:
  - Goodfellow, ch 3, Murphy, ch 2
  - Lecture slides lecture 11, 2018
  - Mitchell, chapter 6