Histograms

As one of the non-parametric techniques in Bayesian Classification

Contents

- Theorem of Bayes
- Bayesian classification
- Non-parametric methods:
  - Histogram: 1-D Case
  - Histogram: 2-D Case
- Example
- Naive Bayes Model
- Discussion

Theorem of Bayes and Image Analysis

- The task of image analysis is to get an explicit description of objects in the image
- This requires to detect objects in the first place
- Therefore, knowledge about the appearance of objects in the image is used
- According to the way the knowledge is represented, there are:
  - model-based methods for image analysis
  - **statistical methods for image analysis (discussed here)**

Theorem of Bayes and Statistical properties

Where does the idea of statistical methods lead us to:

- Objects are not primarily described by object-models, but by statistical properties of the sensor data in relation to the objects
- We need a model of statistic properties in order to recognize objects, this process can be treated as classification
- Observed features can be treated as functions of the object type / class label
- These functions can be represented as probability densities

Theorem of Bayes

Recapitulation of the Theorem of Bayes:

- For the joint distribution $p(\mathbf{x}, C)$, the product rule applies:

$$p(\mathbf{x},C) = p(C \mid \mathbf{x}) \cdot p(\mathbf{x})$$

- Likewise: $p(C, \mathbf{x}) = p(\mathbf{x} \mid C) \cdot p(C)$
- Due to $p(\mathbf{x}, C) = p(C, \mathbf{x})$:

$$p(C \mid \mathbf{x}) \cdot p(\mathbf{x}) = p(x \mid C) \cdot p(C)$$

- Therefore: **Theorem of Bayes:**

$$p(C \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C) \cdot p(C)}{p(\mathbf{x})} = \frac{p(\mathbf{x}, C)}{p(\mathbf{x})}$$

Theorem of Bayes: Interpretation

- C can be treated as object type or class label,
- $\mathbf{x}$ is the observed feature
- $p(C \mid \mathbf{x})$ is posterior probability, a conditional probability for the class label C given the observation $\mathbf{x}$
- $p(\mathbf{x} \mid C)$ is likelihood, the conditional probability to observe a feature given a class
- $p(C)$ corresponds to prior for the occurrence of class label C
- $p(\mathbf{x})$ is probability of the data, the marginal distribution of $\mathbf{x}$, enables us to interpret $p(C \mid \mathbf{x})$ as a probability
- $p(C, \mathbf{x})$ is joint distribution

Theorem of Bayes for classification

- Maximum a posteriori (MAP) criterion: class label C is determined so that the conditional probability $p(C \mid \mathbf{x})$ for the class label C given the observed data $\mathbf{x}$ is maximized
- Given:
  - Models for the likelihoods $p(\mathbf{x} \mid C = L^k)$ of all classes L^k
  - Priori probabilities $p(C = L^k)$ of all classes L^k
  - A feature vector $\mathbf{x}$ to be classified
- Wanted:
  - class C_{map} of $\mathbf{x}$ according to the MAP criterion

Bayesian Classification

- Posterior probability needs to be modeled but it is difficult to be modelled directly
- Instead it can be modelled indirectly using inverse reasoning, which means to derive information about the cause (the object type) from the effect (the observed features)

  → Bayesian Classification

- MAP can also be applied without knowing $p(\mathbf{x})$, since

  $$p(C \mid \mathbf{x}) \propto p(\mathbf{x} \mid C) \cdot p(C)$$

  implies that $\max(p(C \mid \mathbf{x})) = \max(p(\mathbf{x} \mid C) \cdot p(C))$

Bayesian Classification

- Procedure:

  1) For all classes L^k: calculate $p(\mathbf{x}, C=L^k) = p(\mathbf{x}|C=L^k) \cdot p(C=L^k)$
  2) Calculate $p(\mathbf{x}) = \sum_k p(\mathbf{x} \mid C = L^k) \cdot p(C = L^k)$
  3) For all classes L^k : calculate $p(C=L^k \mid \mathbf{x}) = p(\mathbf{x}, C=L^k) / p(\mathbf{x})$
  4) C_{map} is the label of L^k for which $p(C=L^k \mid \mathbf{x})$ is a maximum

- Next step:

  – model $p(\mathbf{x} \mid C)$ directly from the training data:
    Histograms, as an example of non-parametric method

Histograms: 1-D Case

- For the case of discrete variables: $p\left(x = g \mid C = L^k\right) = \frac{K_{gk}}{N_k}$

  K_{gk}: Number of pixels in the training area of class L^k and grey value g;
  N_k: Number of pixels in the training area of class L^k

  Practically implemented with lookup tables!

- For the case of continuous variables: $p(\mathbf{x} = g \mid C = L^k) = \frac{K_{gk}}{N_k \cdot \Delta}$

  Δ: Grid width used for discretization e.g. determined from cross-validation.
  too small value leads to noisy approximation,
  too large leads to strong smoothing!

Histograms: 2-D Case

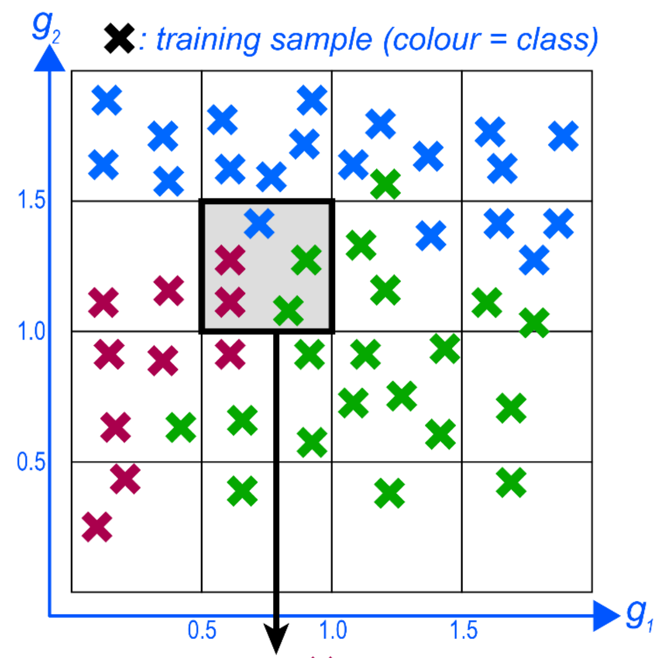
- In case of 2 dimensional discrete variables (assuming $\Delta_1 = \Delta_2$):

$$p\left(x_1 = g_1, x_2 = g_2 \mid C = L^k\right) = \frac{K_{g_1 g_2 k}}{N \cdot \Delta^2}$$

$$= \frac{\text{no. of pixels with class } L^k \text{ with grey value combination } (g_1, g_2)}{\text{no. of pixels with class } L^k \text{ times grid size } \Delta^2}$$

Example

- Image primitives represented as 2-D features
- When g_1 is fixed, there are 4 different options for g_2.
- Also 4 different options, when g_2 is fixed.
- In the end, 16 likelihoods need to be calculated for this 2-D feature space and grid size of 0.5.



$p(1.0 > g_{1,2} \geq 0.5 \mid C = ✖) = 2 / (10 \cdot 0.5^2) = 0.8$

$p(1.0 > g_{1,2} \geq 0.5 \mid C = ✖) = 1 / (20 \cdot 0.5^2) = 0.2$

$p(1.0 > g_{1,2} \geq 0.5 \mid C = ✖) = 2 / (20 \cdot 0.5^2) = 0.4$

Histograms: 2-D Case

- In general Q^D probabilities need to be determined, when we have D dimensional features with Q possible values.

→ Hardly possible for $D > 2$!

→ „Curse of dimensionality“

→ „Hughes phenomenon“ [Hughes, 1968 (!)]:
Beyond a certain point, the classification accuracy is reduced by using additional features

Histograms: 2-D Case

- Can the problem be simplified by determination of the probabilities for each feature **independently**?

the joint distribution of two features x_1, x_2 → $p(x_1, x_2, C) = p(x_1, x_2 \mid C) \cdot p(C)$

→ Generally not possible, but..

- If we assume the two features x_1, x_2 to be conditionally independent, we can factorize the likelihood $p(x_1, x_2 \mid C)$ to $p(x_1, x_2 \mid C) = p(x_1 \mid C) \cdot p(x_2 \mid C)$
- By definition, the features x_1 and x_2 are conditionally independent if $p(x_1 \mid x_2, C)$ does **not** depend on x_2

Histograms: 2-D Case

- "conditionally independent" thus means that x_1 and x_2 are statistically independent if C is known. It does not mean that x_1 and x_2 are statistically independent in the general meaning of the word.

- We can extend it, if the features of a multi-dimensional feature vector **x** are conditionally independent, the likelihood can be factorised: $p(\boldsymbol{x} \mid C) = p(\boldsymbol{x}_1 \mid C) \cdot p(\boldsymbol{x}_2 \mid C) \cdot \cdot\, p(\boldsymbol{x}_D \mid C)$

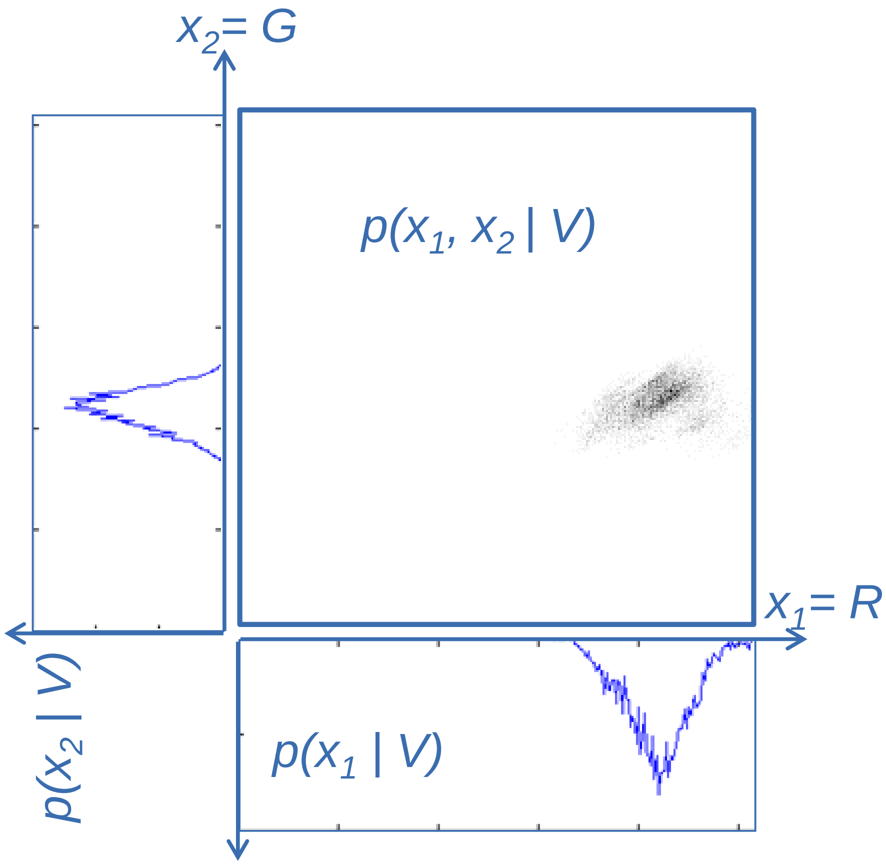
- Consequence: the likelihood can be determined from the marginal distributions $p(\boldsymbol{x}_i \mid C)$

→ $Q \cdot D$ instead of Q^D parameters!

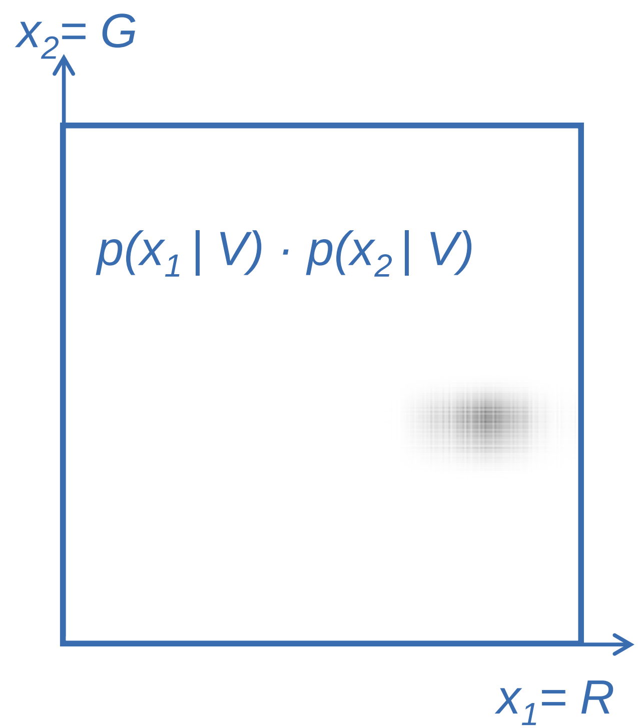
- This approach is called the „Naive Bayes Model“

Example of Impact of the Naive Bayes Model

- Aerial image with training area for “*vegetation*“ (V)
  (87 x 85 = 7395 pixels)



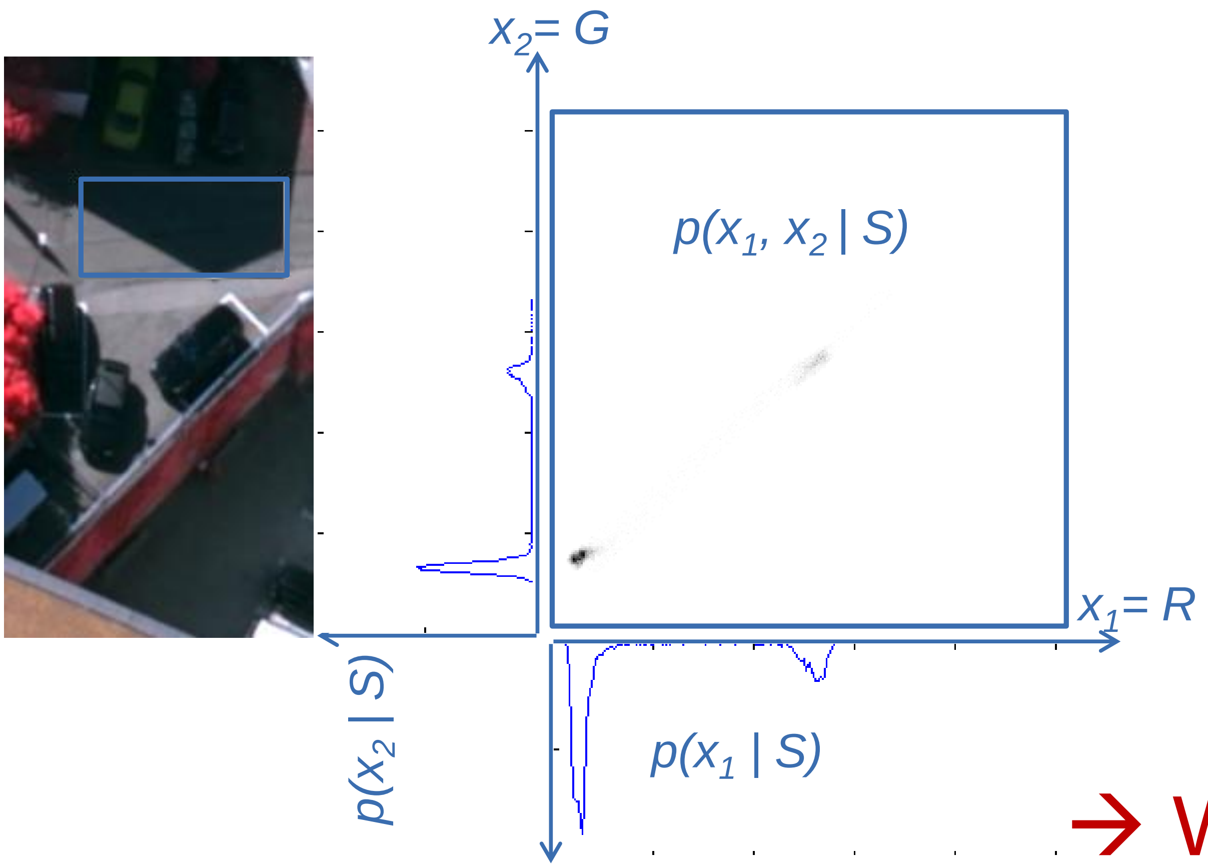
Assuming conditional independence:



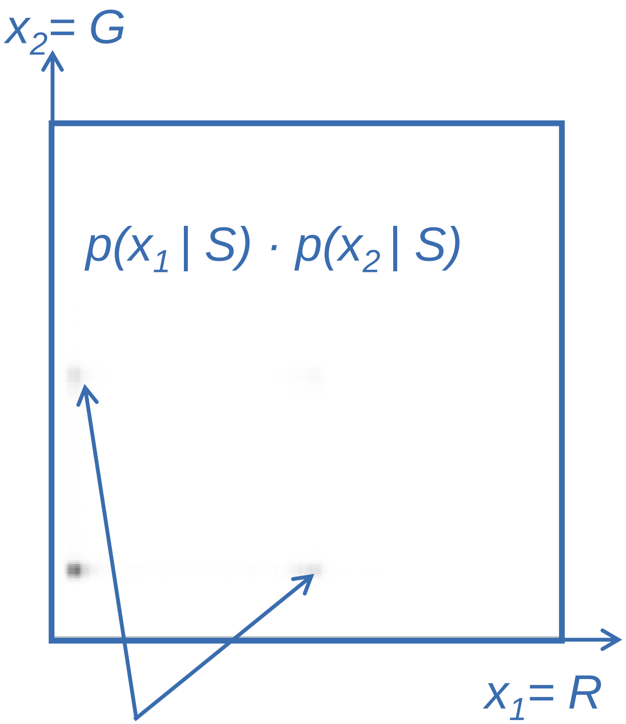
→ MODEL OK!

Example of Impact of the Naive Bayes Model

Aerial image with training area for “*street*“ (S)
(49 x 102 = 4998 pixels)



Assuming conditional independence:



incorrect clusters!

→ WRONG MODEL!

Discussion

- Bayesian classification uses „inverse reasoning“ since likelihoods are often easier to model than posteriors
- Using histograms as a non-parametric technique to model the likelihoods is a simple but often well working approach
- Histograms can also be used for multidimensional data, but for more than two dimensions the amount of required training data and computational resources drastically increases

  → Curse of Dimensionality

Discussion

- One way to still use non-parametric techniques for multidimensional data is the Naive Bayes Model
- In the Naive Bayes Model statistical dependencies between the features are neglected, which is a strong simplification in general! Maybe can be justified if the features are determined from independent sensors
- However, wrongly taking the assumption of conditional independence can lead to a incorrect likelihood model