

Classifier Combination

Computational Social Intelligence - Lecture 20

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

<http://www.dcs.gla.ac.uk/vincia>
Alessandro.Vinciarelli@glasgow.ac.uk



University
of Glasgow

EPSRC

Engineering and Physical Sciences
Research Council



This lecture is based on the following text (available on Moodle):

- Vinciarelli & Esposito, “Multimodal Analysis of Social Signals”, in “The Handbook of Multimodal-Multisensor Interfaces”, Oviatt et al. (eds.), 203-226, ACM, 2018

Outline

- Quick Recap
- Late Fusion (Sum Rule)
- Variants of Late Fusion
- Conclusion

Outline

- Quick Recap
- Late Fusion (Sum Rule)
- Variants of Late Fusion
- Conclusion

There are no changes
for the priors (they do
not depend on the input
vectors)

$$\mathcal{C}^* = \arg \max_{\mathcal{C}_k \in \mathcal{C}} p(\vec{x}_1, \dots, \vec{x}_R | \mathcal{C}_k) p(\mathcal{C}_k)$$

The likelihood must be
changed to reflect the
presence of multiple
feature vectors

The assumption is that the input vectors are statistically independent given the class

Product over all sensors

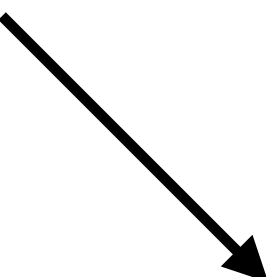
$$p(\vec{x}_1, \dots, \vec{x}_R | \mathcal{C}_k) = \prod_{j=1}^R p(\vec{x}_j | \mathcal{C}_k)$$

If one term is close to zero, the entire product is close to zero

Outline


- Quick Recap
- **Late Fusion (Sum Rule)**
- Variants of Late Fusion
- Conclusion

The Bayes Theorem


$$p(\vec{x}_j | \mathcal{C}_k) = \frac{p(\mathcal{C}_k | \vec{x}_j) p(\vec{x}_j)}{p(\mathcal{C}_k)}$$

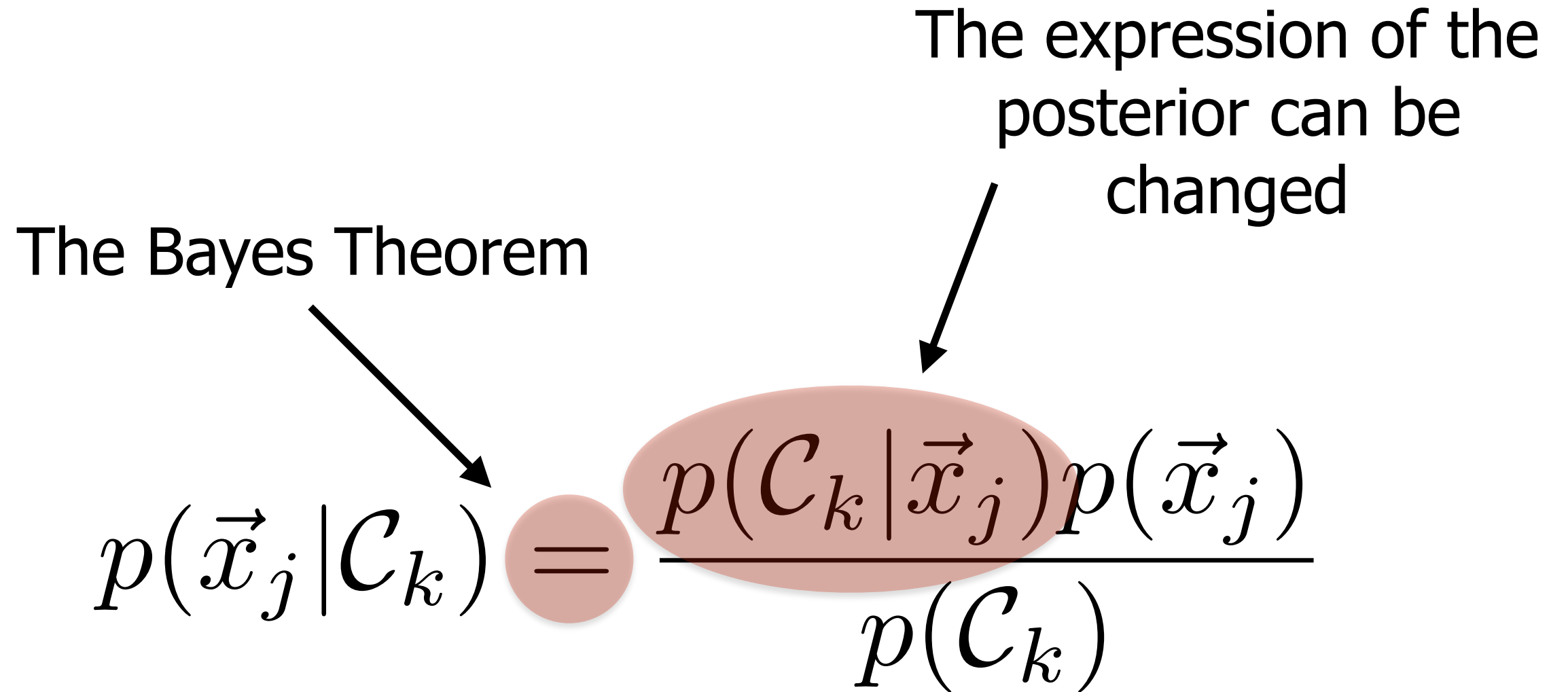
The posterior is
assumed to
approximate the prior

The absolute value is
significantly smaller
than one


$$p(\mathcal{C}_k | \vec{x}_j) \simeq p(\mathcal{C}_k) (1 + \delta_{jk})$$

The Bayes Theorem

The expression of the
posterior can be
changed



The diagram shows the Bayes' Theorem equation with two annotations. An arrow points from the text 'The Bayes Theorem' to the left side of the equation, $p(\vec{x}_j | \mathcal{C}_k)$. Another arrow points from the text 'The expression of the posterior can be changed' to the fraction $\frac{p(\mathcal{C}_k | \vec{x}_j)p(\vec{x}_j)}{p(\mathcal{C}_k)}$. The fraction is highlighted with a red oval, and the equals sign is highlighted with a red circle.

$$p(\vec{x}_j | \mathcal{C}_k) = \frac{p(\mathcal{C}_k | \vec{x}_j)p(\vec{x}_j)}{p(\mathcal{C}_k)}$$

$$p(\vec{x}_j | \mathcal{C}_k) = \frac{p(\vec{x}_k)(1 + \delta_{jk})p(\vec{x}_j)}{p(\vec{x}_j)}$$

The assumption is that
the input vectors are
statistically independent
given the class

Product over all sensors

The diagram illustrates the joint probability distribution $p(\vec{x}_1, \dots, \vec{x}_R | \mathcal{C}_k)$ as a product over all sensors $j=1$ to R of the individual sensor distributions $p(\vec{x}_j | \mathcal{C}_k)$. The text "The assumption is that the input vectors are statistically independent given the class" points to the left side of the equation. The text "Product over all sensors" points to the product symbol and the index $j=1$ to R . The term $p(\vec{x}_j | \mathcal{C}_k)$ is highlighted with a red oval, and a curved arrow points from the text "If one term is close to zero, the entire product is close to zero" to this term.

$$p(\vec{x}_1, \dots, \vec{x}_R | \mathcal{C}_k) = \prod_{j=1}^R p(\vec{x}_j | \mathcal{C}_k)$$

If one term is close to
zero, the entire product
is close to zero

$$p(\vec{x}_1, \dots, \vec{x}_R | \mathcal{C}_k) = \prod_{j=1}^R (1 + \delta_{jk}) p(\vec{x}_j)$$

The factors of the product are rearranged

This term does not depend on the class

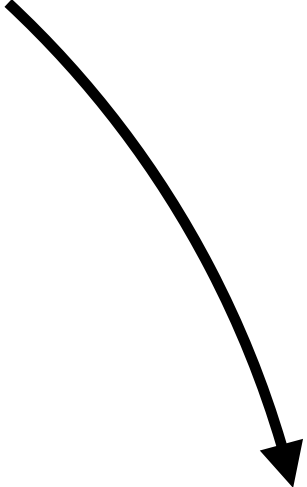
$$\prod_{j=1}^R (1 + \delta_{jk}) p(\vec{x}_j) = \prod_{j=1}^R (1 + \delta_{jk}) \prod_{j=1}^R p(\vec{x}_j)$$

This product can be neglected because it has always the same value for all classes

This term can be
neglected because the
delta's are small


$$\prod_{j=1}^R (1 + \delta_{jk}) = (1 + \delta_{1k} + \delta_{2k} + \delta_{1k}\delta_{2k}) \prod_{j=3}^R (1 + \delta_{jk}) \simeq (1 + \delta_{1k} + \delta_{2k}) \prod_{j=3}^R (1 + \delta_{jk})$$

The terms of the product that include several delta's can be neglected


$$\prod_{j=1}^R (1 + \delta_{jk}) \approx 1 + \sum_{j=1}^R \delta_{jk}$$

The posterior is
assumed to
approximate the prior

The absolute value is
significantly smaller
than one

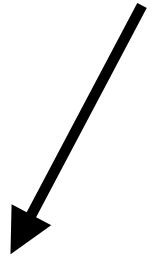


The diagram features two arrows. The first arrow originates from the text 'The posterior is assumed to approximate the prior' and points to the approximation symbol \simeq in the equation. The second arrow originates from the text 'The absolute value is significantly smaller than one' and points to the term δ_{jk} in the equation.

$$p(\mathcal{C}_k | \vec{x}_j) \simeq p(\mathcal{C}_k) (1 + \delta_{jk})$$

$$\delta_{jk} = \frac{p(\mathcal{C}_k | \vec{x}_j)}{p(\mathcal{C}_k)} - 1$$

The expression of the
delta's is replaced in the
product of the
likelihoods


$$\prod_{j=1}^R p(\vec{x}_j | \mathcal{C}_k) = 1 + \sum_{j=1}^R \left[\frac{p(\vec{C}_k | \vec{x}_j)}{p\mathcal{C}_k} - 1 \right]$$

There are no changes
for the priors (they do
not depend on the input
vectors)

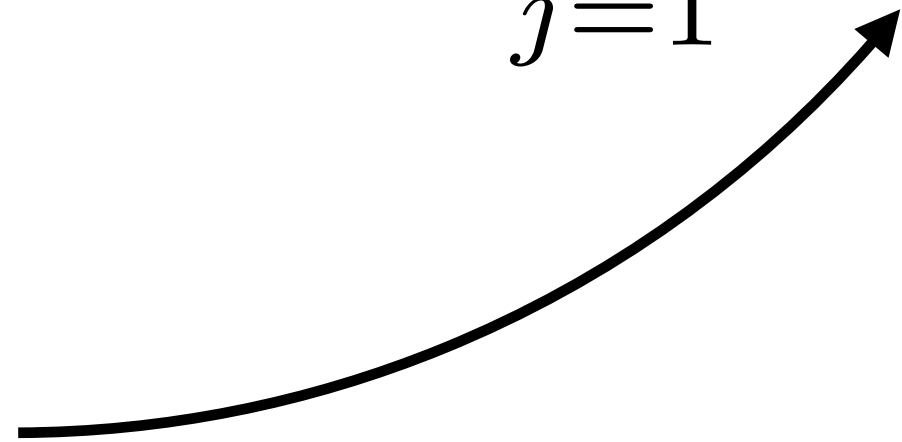
$$\mathcal{C}^* = \arg \max_{\mathcal{C}_k \in \mathcal{C}} p(\vec{x}_1, \dots, \vec{x}_R | \mathcal{C}_k) p(\mathcal{C}_k)$$

The likelihood must be
changed to reflect the
presence of multiple
feature vectors

The Sum Rule

$$\mathcal{C}^* = \arg \max_k (1 - R)p(\mathcal{C}_k) + \sum_{j=1}^R p(\mathcal{C}_k | \vec{x}_j)$$

The sum of the
posteriors when using
the individual modalities



Outline

- Quick Recap
- Late Fusion (Sum Rule)
- **Variants of Late Fusion**
- Conclusion

The product of the posteriors is bound by the minimum of the posteriors

$$\prod_{j=1}^R p(\mathcal{C}_k | \vec{x}_j) \leq \min_j p(\mathcal{C}_k | \vec{x}_j) \leq \frac{1}{R} \sum_{j=1}^R p(\mathcal{C}_k | \vec{x}_j) \leq \max_k p(\mathcal{C}_k | \vec{x}_j)$$

The sum of the posteriors is bound by the maximum of the posteriors

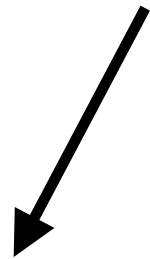
$$\mathcal{C}^* = \arg \max_k (1 - R)p(\mathcal{C}_k) + \sum_{j=1}^R p(\mathcal{C}_k | \vec{x}_j)$$

$$= \arg \max_k (1 - R)p(\mathcal{C}_k) + R \max_k p(\mathcal{C}_k | \vec{x}_j)$$

The Max Rule



When the priors are
uninformative



$$\mathcal{C}^* = \arg \max_{jk} p(\mathcal{C}_k | \vec{x}_j)$$

Max Rule when the
priors are uninformative

There are no changes
for the priors (they do
not depend on the input
vectors)

$$\mathcal{C}^* = \arg \max_{\mathcal{C}_k \in \mathcal{C}} p(\vec{x}_1, \dots, \vec{x}_R | \mathcal{C}_k) p(\mathcal{C}_k)$$

The likelihood must be
changed to reflect the
presence of multiple
feature vectors

$$\mathcal{C}^* = \arg \max_k p(\mathcal{C}_k) \prod_{j=1}^R p(\vec{x}_j | \mathcal{C}_k)$$

The likelihood is
rewritten using the
Bayes Theorem

The Min Rule

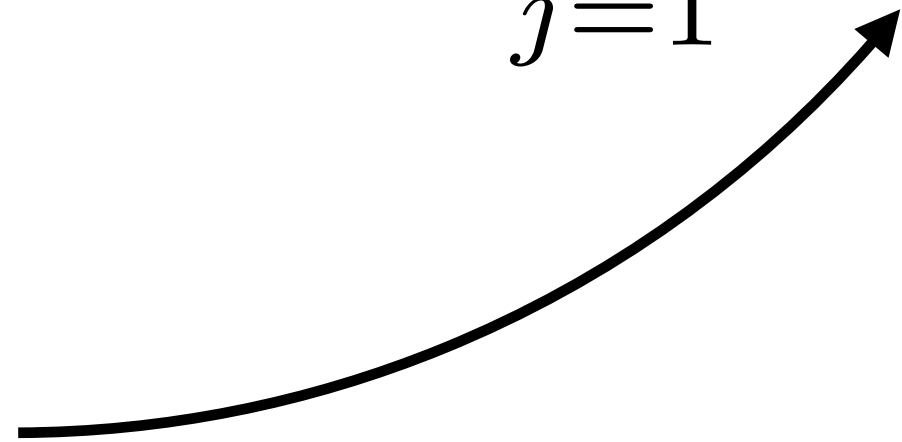
$$\mathcal{C}^* = \arg \max_k \prod_{j=1}^R p(\mathcal{C}_k | \vec{x}_j) =$$
$$= \arg \max_k \min_j p(\mathcal{C}_k | \vec{x}_j)$$

The class where the
minimum of the
posteriors is the highest

The Sum Rule

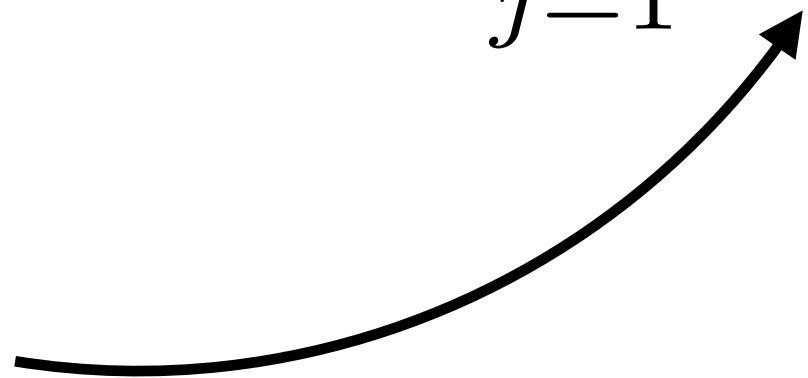
$$\mathcal{C}^* = \arg \max_k (1 - R)p(\mathcal{C}_k) + \sum_{j=1}^R p(\mathcal{C}_k | \vec{x}_j)$$

The sum of the
posteriors when using
the individual modalities



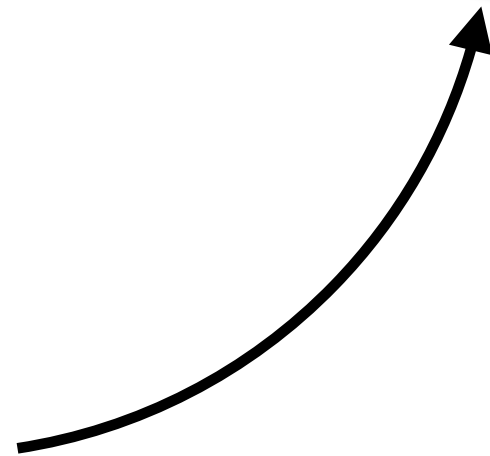
$$\mathcal{C}^* = \arg \max_k (1 - R)p(\mathcal{C}_k) + \frac{1}{R} \sum_{j=1}^R p(\mathcal{C}_k | \vec{x}_j)$$

The average can be
noisy when R is small



$$\mathcal{C}^* = \arg \max_k (1 - R)p(\mathcal{C}_k) + M_j p(\mathcal{C}_k | \vec{x}_j)$$

The median of the
posteriors for a given
class



Outline

- Quick Recap
- Late Fusion (Sum Rule)
- **Variants of Late Fusion**
- Conclusion

Conclusions

- The combination of multiple classifiers is the methodology underlying multimodal approaches;
- The early fusion works when the number of feature vectors is the same across multiple modalities;
- The late fusion works when the number of feature vectors is different for different modalities.