Discriminant Functions

Computational Social Intelligence - Lecture 13

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

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

Texts (see Moodle)

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

- Chapter 5 of F.Camastra and A.Vinciarelli, "Machine Learning for Audio, Image and Video Processing", Springer Verlag, 2008.

Outline

- Quick Recap
- Conditional Risk and Bayes Decision Rule
- Binary Classification and Likelihood Ratio
- Zero-One Loss and Posterior Rule
- Gaussian Discriminant Functions
- Conclusions

Outline

- Quick Recap
- Conditional Risk and Bayes Decision Rule
- Binary Classification and Likelihood Ratio
- Zero-One Loss and Posterior Rule
- Gaussian Discriminant Functions
- Conclusions

The Bayes Theorem

A-posteriori probability (or posterior) of class i

Likelihood of class i with respect to the feature vector

$$p(C_i|\vec{x}) = \frac{p(\vec{x}|C_i)p(C_i)}{p(\vec{x})}$$

The evidence

The a-priori probability of class i

Error Probability

Probability of error if the right decision is j

$$p(err|\vec{x}) = \sum_{i \neq j} p(C_i|\vec{x})$$

The sum over all posteriors except the posterior of j

Error Probability

The class that corresponds to the highest posterior

$$C^* = \arg\max_{C_k \in C} p(C_k|\vec{x})$$

The value of the posterior is checked for all possible classes

The posterior takes the features into account

Posterior Rule

The expression of the priors according to the Bayes Theorem

$$C^* = \arg\max_{C_k \in C} \frac{p(\vec{x}|C_k)p(C_k)}{p(\vec{x})} =$$

$$= \arg\max_{C_k \in C} p(\vec{x}|C_k)p(C_k)$$

The evidence is the same for all classes and it can be eliminated

Outline

- Quick Recap
- Conditional Risk and Bayes Decision Rule
- Binary Classification and Likelihood Ratio
- Zero-One Loss or Posterior Rule
- Gaussian Discriminant Functions
- Conclusions

A possible decision

A feature vector

$$p(C_j|\vec{x}) \qquad j = 1, \ldots, M$$

A posteriori probability of decision "j"

Number of possible decisions

The loss resulting from performing action "i" when the decision is "j"

A possible action

A possible decision

Number of possible actions

$$\pi(\mathcal{B}_i|\mathcal{C}_j)$$

$$i = 1, \ldots, T$$

$$j = 1, \ldots, M$$

Number of possible decisions

Expected loss (or conditional risk) associated to the action

A possible action

$$\mathcal{R}(\mathcal{B}_i|\vec{x}) = \sum_{j=1}^{M} \pi(\mathcal{B}_i|\mathcal{C}_j) p(\mathcal{C}_j|\vec{x})$$

Sum over all possible classes the vector can be assigned to

The posterior of the class acts as a weight in the sum

Bayes Decision Rule

The index of the action that minimises the conditional risk

$$\hat{k} = \arg \min_{k=1,\ldots,T} \mathcal{R}(\mathcal{B}_k|\vec{x}) =$$

$$\mathcal{R}(\mathcal{B}_{\hat{k}}|\vec{x})$$

The Bayes Risk

Recap

- The conditional risk is the weighted sum of the losses associated to an action;
- The weights are the posteriors of the classes (the conditional risk is an expectation);
- The Bayes Decision Rule targets the decision that minimises the conditional risk.

Outline

- Quick Recap
- Conditional Risk and Bayes Decision Rule
- Binary Classification and Likelihood Ratio
- Zero-One Loss and Posterior Rule
- Gaussian Discriminant Functions
- Conclusions

Binary Classification

- In a binary classification there are two classes and two actions;
- Each of the two actions is right for one class, but wrong for the other;
- For example, the classes are “healthy” and “ill” while the actions are “keep in the hospital” and “do not keep in the hospital”.

Expected loss (or conditional risk) associated to action 1

$$\mathcal{R}(\mathcal{B}_1|\vec{x}) = \pi(\mathcal{B}_1|\mathcal{C}_1)p(\mathcal{C}_1|\vec{x}) + \pi(\mathcal{B}_1|\mathcal{C}_2)p(\mathcal{C}_2|\vec{x})$$

$$\mathcal{R}(\mathcal{B}_2|\vec{x}) = \pi(\mathcal{B}_2|\mathcal{C}_1)p(\mathcal{C}_1|\vec{x}) + \pi(\mathcal{B}_2|\mathcal{C}_2)p(\mathcal{C}_2|\vec{x})$$

Expected loss (or conditional risk) associated to action 2

When the difference is positive, action 1 minimises the expected loss

$$\mathcal{R}(\mathcal{B}_2|\vec{x}) - \mathcal{R}(\mathcal{B}_1|\vec{x}) =$$

$$= [\pi(\mathcal{B}_2|\mathcal{C}_1) - \pi(\mathcal{B}_1|\mathcal{C}_1)]\, p(\mathcal{C}_1|\vec{x}) +$$

$$+ [\pi(\mathcal{B}_2|\mathcal{C}_2) - \pi(\mathcal{B}_1|\mathcal{C}_2)]\, p(\mathcal{C}_2|\vec{x})$$

The Likelihood Ratio

$$\frac{p(\vec{x}|\mathcal{C}_1)}{p(\vec{x}|\mathcal{C}_2)} > \frac{\pi(\mathcal{B}_1|\mathcal{C}_2) - \pi(\mathcal{B}_2|\mathcal{C}_2)}{\pi(\mathcal{B}_2|\mathcal{C}_1) - \pi(\mathcal{B}_1|\mathcal{C}_1)} \frac{p(\mathcal{C}_2)}{p(\mathcal{C}_1)}$$

When the Likelihood Ratio satisfies the equation above, action 1 minimises the expected loss

Example

- The authentication of a face image for unlocking a device is an example of binary classification;
- Decision 1 is to accept the face image (the action is unlock), Decision 2 is to reject the face (the action is to keep the device locked);
- The faces are converted into feature vectors through image processing techniques.

Likelihood of accepting

Loss resulting from unlocking when the right decision is reject

Loss resulting from locking when the right decision is reject

$$\frac{p(\vec{x}|\mathcal{C}_a)}{p(\vec{x}|\mathcal{C}_r)} \frac{\pi(u|r) - \pi(l|r)}{\pi(l|a) - \pi(u|a)}$$

Likelihood of rejecting

Loss resulting from locking when the right decision is accept

Loss resulting from unlocking when the right decision is accept

Likelihood of accepting

Loss resulting from unlocking when the right decision is reject

Loss resulting from locking when the right decision is reject

$$\frac{p(\vec{x}|\mathcal{C}_a)}{p(\vec{x}|\mathcal{C}_r)} = \frac{10 - 0}{5 - 0} = 2$$

Likelihood of rejecting

Loss resulting from locking when the right decision is accept

Loss resulting from unlocking when the right decision is accept

Recap

- A binary decision problem can be addressed through the likelihood ratio;
- The losses and the priors determine the likelihood ratio threshold above which one of the two decisions is made;

Outline

- Quick Recap
- Conditional Risk and Bayes Decision Rule
- Binary Classification and Likelihood Ratio
- Zero-One Loss and Posterior Rule
- Gaussian Discriminant Functions
- Conclusions

The loss resulting from performing action "i" when the decision is "j"

There is no loss resulting from performing action "i" when the decision is "i"

$$\pi(\mathcal{B}_i|\mathcal{C}_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

The loss is the same for every mismatch between decision and action

The expected loss (conditional risk) when performing action "i"

The sum includes only the cases in which "j" and "i" are different

$$\mathcal{R}(\mathcal{B}_i|\vec{x}) = \sum_{j \neq i} \pi(\mathcal{B}_i|\mathcal{C}_j) p(\mathcal{C}_j|\vec{x}) =$$

$$= 1 - p(\vec{C}_i|\vec{x})$$

The highest posterior minimises the conditional risk

Recap

- In most cases, the number of actions is the same as the number of decisions;
- In general, for every class one decision is right while all the others are wrong;
- In such cases, the minimisation of the conditional risk is equivalent to the application of the posterior rule.

Outline

- Quick Recap
- Conditional Risk and Bayes Decision Rule
- Binary Classification and Likelihood Ratio
- Zero-One Loss and Posterior Rule
- Gaussian Discriminant Functions
- Conclusions

The index of the function gamma that corresponds to the maximum

All functions in set G

$$\hat{k} = \arg \max_{k=1,\ldots,M} \gamma_k(\vec{x})$$

$$\mathcal{G} = \{\gamma_1(\vec{x}), \ldots, \gamma_M(\vec{x})\}$$

The discriminant functions

The discriminant functions are as many as the classes

The opposite of the expected loss can be used as discriminant function

This relationship holds for a zero-one loss function

$$\gamma_k(\vec{x}) = -\mathcal{R}(\mathcal{B}_k|\vec{x}) = p(\mathcal{C}_k|\vec{x}) - 1$$

$$\gamma_k(\vec{x}) \simeq p(\mathcal{C}_k|\vec{x})$$

The discriminant functions are compared to one another, the additive constant can be dropped

The posterior of decision “k”

This relationship holds thanks to the Bayes Theorem

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

The logarithm is monotonic, the result of the comparison does not change

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

$$\log \gamma_k(\vec{x}) \simeq \log p(\vec{x}|\mathcal{C}_k) + \log p(\mathcal{C}_k)$$

The evidence is the same for all discriminant functions and can be dropped

The decision corresponds to the maximum value of the logarithm

$$\hat{k} = \arg \max_{k \in [1,M]} \log p(\vec{x}|\mathcal{C}_k) + \log p(\mathcal{C}_k)$$

All possible values of k are tested

The Likelihood

The probability of a vector is the joint probability of its components

$$p(\vec{x}|\mathcal{C}_k) = p(x_1, x_2, \ldots, x_D|\mathcal{C}_k)$$

The dimensionality of the vector

The probability of a vector is the joint probability of its components

Likelihood of an individual component

$$p(x_1, x_2, \ldots, x_D|\mathcal{C}_k) = \prod_{i=1}^{D} p(x_i|\mathcal{C}_k)$$

It is true if the components are statistically independent given the class

The Naive Bayes classifier

The decision corresponds to the maximum value of the logarithm

$$\hat{k} = \arg \max_{k \in [1,M]} \log p(\vec{x}|\mathcal{C}_k) + \log p(\mathcal{C}_k)$$

All possible values of k are tested

$$\log p(\vec{x}|\mathcal{C}_k) = \sum_{i=1}^{D} \log p(x_i|\mathcal{C}_k)$$

It is true if the components are statistically independent given the class

Probability of observing the value of feature "i" when the class is "k"

Average of feature "i" in class "k"

$$p(x_i|C_k) = \frac{1}{\sqrt{2\pi}\sigma_{ik}} \exp\left[-\frac{(x_i - \mu_{ik})^2}{2\sigma_{ik}^2}\right]$$

Standard deviation of feature "i" in class "k"

Sum over the standard deviations of the individual features

True when features statistically independent given the class

$$\log p(\vec{x}|\mathcal{C}_k) = -\sum_{i=1}^{D}\left[\log\sqrt{2\pi\sigma_{ik}} + \frac{(x_i - \mu_{ik})^2}{2\sigma_{ik}^2}\right]$$

Euclidean distance between feature vector and class average

Outline

- Quick Recap
- Conditional Risk and Bayes Decision Rule
- Binary Classification and Likelihood Ratio
- Zero-One Loss and Posterior Rule
- Gaussian Discriminant Functions
- Conclusions

Conclusions

- Discriminant functions allow one to implement the Bayes Decision Rule by checking their value for the feature vectors;
- The features can be assumed statistically independent given the class (the Naive Bayes Classifier);
- The discriminant functions can be thought of distances from the class averages.