7b - Feature Selection II

COM2004/3004

Jon Barker

Department of Computer Science
University of Sheffield

Autumn Semester

Lecture Objectives

In this lecture we will,

- Consider the problem of selecting **multiple** features.
- Explain why divergence of **1-D distributions** is not generally useful.
- Examine divergence for multivariate normal distributions.
- Explore how feature selection can be automated using a selection algorithm.
- Compare various feature selection algorithms.

Recap

- Feature selection is about choosing features that minimise the probability of classification error.
- Classification error arises when class distributions 'overlap'.
- So we choose features whose distributions have the smallest 'overlap'.
- We talk about 'measures of *class separability*'.
- We introduced 'Divergence', d_{12}, as one such measure.

Recap

- Feature selection is about choosing features that minimise the probability of classification error.
- Classification error arises when class distributions 'overlap'.
- So we choose features whose distributions have the smallest 'overlap'.
- We talk about 'measures of *class separability*'.
- We introduced 'Divergence', d_{12}, as one such measure.

Recap

- Feature selection is about choosing features that minimise the probability of classification error.
- Classification error arises when class distributions ‘overlap’.
- So we choose features whose distributions have the smallest ‘overlap’.
- We talk about ‘measures of *class separability*’.
- We introduced ‘Divergence’, d_{12}, as one such measure.

Recap

- Feature selection is about choosing features that minimise the probability of classification error.
- Classification error arises when class distributions ‘overlap’.
- So we choose features whose distributions have the smallest ‘overlap’.
- We talk about ‘measures of *class separability*’.
- We introduced ‘Divergence’, d_{12}, as one such measure.

Recap

- Feature selection is about choosing features that minimise the probability of classification error.
- Classification error arises when class distributions ‘overlap’.
- So we choose features whose distributions have the smallest ‘overlap’.
- We talk about ‘measures of *class separability*’.
- We introduced ‘Divergence’, d_{12}, as one such measure.

Divergence

- Define,

$$D_{12} = \int_{-\infty}^{+\infty} p(x|\omega_1) \ln \frac{p(x|\omega_1)}{p(x|\omega_2)} dx \quad (1)$$

- The **divergence**, written d_{12}, is defined as,

$$d_{12} = D_{12} + D_{21} \quad (2)$$

- For 1-d Gaussians the equation for divergence becomes,

$$d_{12} = \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2} - 2\right) + \frac{1}{2}(\mu_1 - \mu_2)^2 \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)$$

where μ_1 and μ_2 are the means of the distributions, and σ_1 and σ_2 are the standard deviations.

Divergence

- Define,

$$D_{12} = \int_{-\infty}^{+\infty} p(x|\omega_1) \ln \frac{p(x|\omega_1)}{p(x|\omega_2)} dx \qquad (1)$$

- The **divergence**, written d_{12}, is defined as,

$$d_{12} = D_{12} + D_{21} \qquad (2)$$

- For 1-d Gaussians the equation for divergence becomes,

$$d_{12} = \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2} - 2\right) + \frac{1}{2}(\mu_1 - \mu_2)^2 \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)$$

where μ_1 and μ_2 are the means of the distributions, and σ_1 and σ_2 are the standard deviations.

Divergence

- Define,

$$D_{12} = \int_{-\infty}^{+\infty} p(x|\omega_1) \ln \frac{p(x|\omega_1)}{p(x|\omega_2)} dx \tag{1}$$

- The **divergence**, written d_{12}, is defined as,

$$d_{12} = D_{12} + D_{21} \tag{2}$$

- For 1-d Gaussians the equation for divergence becomes,

$$d_{12} = \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2} - 2\right) + \frac{1}{2}(\mu_1 - \mu_2)^2 \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)$$

where μ_1 and μ_2 are the means of the distributions, and σ_1 and σ_2 are the standard deviations.

Selecting features

- Consider again our gender features.
- Say we have computed the divergence between male and female distributions for each feature:

left arm length	2.5
right arm length	2.4
hair length	1.8
weight	1.3
shoe size	1.2
eye colour	0.01[1]

- If using one feature, best feature is '**left arm length**'
- If using two features, best features would be '**left arm length**' and '**right arm length**'... Right?
- **No!**, not necessarily... we need to consider their joint distribution.

[1] These figures are made up for sake of illustration and are not intended to be realistic!

Selecting features

- Consider again our gender features.
- Say we have computed the divergence between male and female distributions for each feature:

left arm length	2.5
right arm length	2.4
hair length	1.8
weight	1.3
shoe size	1.2
eye colour	0.01[1]

- If using one feature, best feature is '**left arm length**'
- If using two features, best features would be '**left arm length**' and '**right arm length**'... Right?
- **No!**, not necessarily...we need to consider their joint distribution.

[1] These figures are made up for sake of illustration and are not intended to be realistic!

Selecting features

- Consider again our gender features.
- Say we have computed the divergence between male and female distributions for each feature:

left arm length	2.5
right arm length	2.4
hair length	1.8
weight	1.3
shoe size	1.2
eye colour	0.01[1]

- If using one feature, best feature is '**left arm length**'
- If using two features, best features would be '**left arm length**' and '**right arm length**'... Right?
- **No!**, not necessarily... we need to consider their joint distribution.

[1] These figures are made up for sake of illustration and are not intended to be realistic!

Selecting features

- Consider again our gender features.
- Say we have computed the divergence between male and female distributions for each feature:

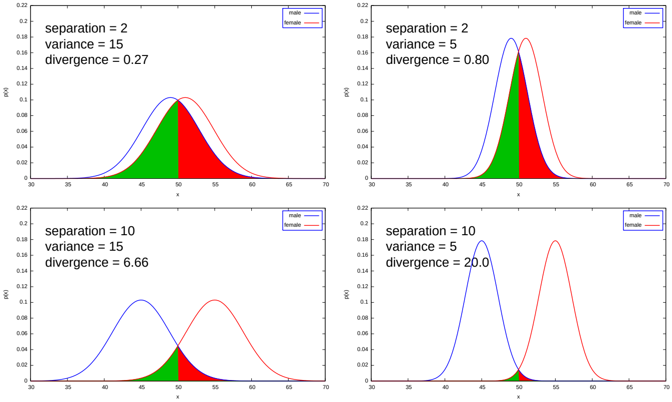
left arm length	2.5
right arm length	2.4
hair length	1.8
weight	1.3
shoe size	1.2
eye colour	0.01[1]

- If using one feature, best feature is '**left arm length**'
- If using two features, best features would be '**left arm length**' and '**right arm length**'... Right?
- **No!**, not necessarily... we need to consider their joint distribution.

[1] These figures are made up for sake of illustration and are not intended to be realistic!

Divergence in 1 dimension

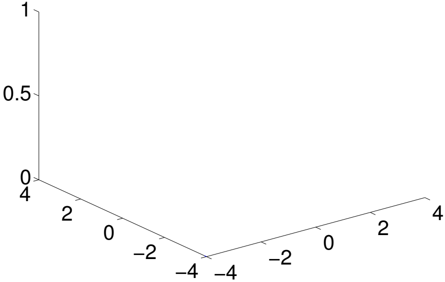
Divergence for pairs of Gaussian distributions:



But distributions have more opportunity to be separate in higher dimensions...

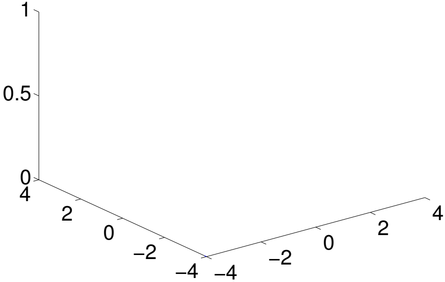
Pairs of features

- Consider the joint distribution of a pair of features, x_1 and x_2 – written as $p(x_1, x_2)$.
- We can plot the value of x_1 on the x-axis and x_2 on the y-axis, and $p(x_1, x_2)$ on the z-axis to form a surface.
- If x_1 and x_2 have a Gaussian distribution *and are independent* then $p(x_1, x_2)$ will look something like this,



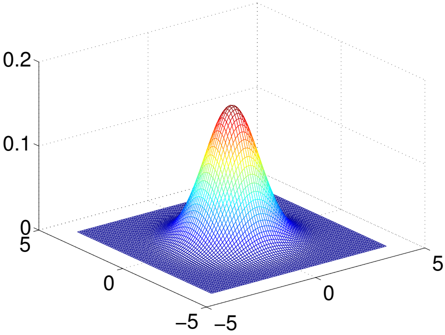
Pairs of features

- Consider the joint distribution of a pair of features, x_1 and x_2 – written as $p(x_1, x_2)$.
- We can plot the value of x_1 on the x-axis and x_2 on the y-axis, and $p(x_1, x_2)$ on the z-axis to form a surface.
- If x_1 and x_2 have a Gaussian distribution *and are independent* then $p(x_1, x_2)$ will look something like this,



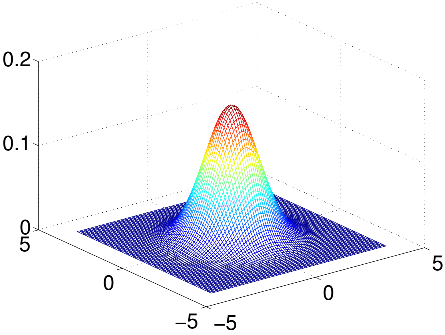
Pairs of features

- Consider the joint distribution of a pair of features, x_1 and x_2 – written as $p(x_1, x_2)$.
- We can plot the value of x_1 on the x-axis and x_2 on the y-axis, and $p(x_1, x_2)$ on the z-axis to form a surface.
- If x_1 and x_2 have a Gaussian distribution *and are independent* then $p(x_1, x_2)$ will look something like this,



Pairs of features

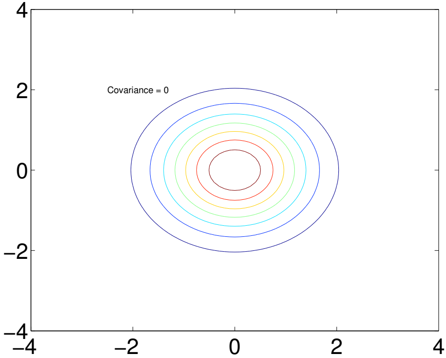
- Consider the joint distribution of a pair of features, x_1 and x_2 – written as $p(x_1, x_2)$.
- We can plot the value of x_1 on the x-axis and x_2 on the y-axis, and $p(x_1, x_2)$ on the z-axis to form a surface.
- If x_1 and x_2 have a Gaussian distribution *and are independent* then $p(x_1, x_2)$ will look something like this,



- ...or with the corresponding contour plot.

Pairs of features

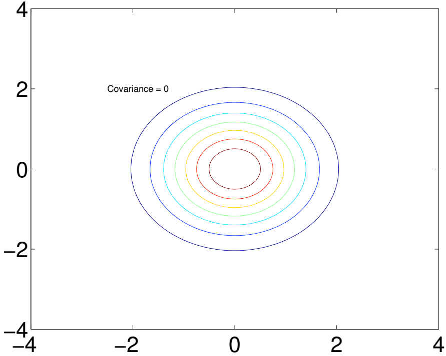
- Consider the joint distribution of a pair of features, x_1 and x_2 – written as $p(x_1, x_2)$.
- We can plot the value of x_1 on the x-axis and x_2 on the y-axis, and $p(x_1, x_2)$ on the z-axis to form a surface.
- If x_1 and x_2 have a Gaussian distribution *and are independent* then $p(x_1, x_2)$ will look something like this,



- . . . or just as a contour plot.

Correlation

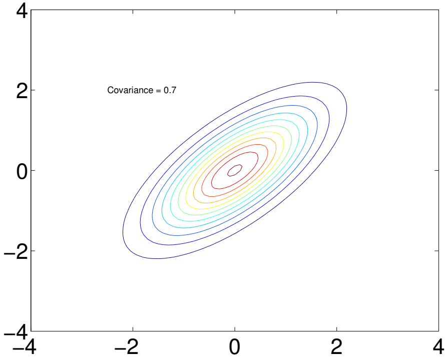
- If x_1 and x_2 are independent and have equal variance the contours of $p(x_1, x_2)$ form circles.
- However, if x_1 and x_2 are correlated the circles become stretched along the diagonal.



- How does this correlation effect the divergence (i.e. overlap) between a pair of distributions.

Correlation

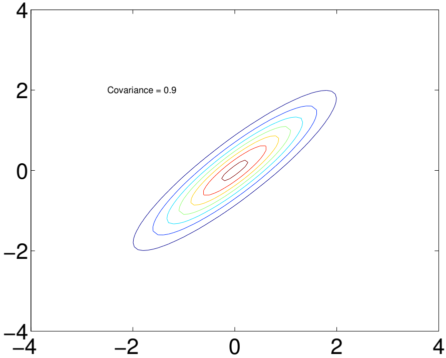
- If x_1 and x_2 are independent and have equal variance the contours of $p(x_1, x_2)$ form circles.
- However, if x_1 and x_2 are correlated the circles become stretched along the diagonal.



- How does this correlation effect the divergence (i.e. overlap) between a pair of distributions.

Correlation

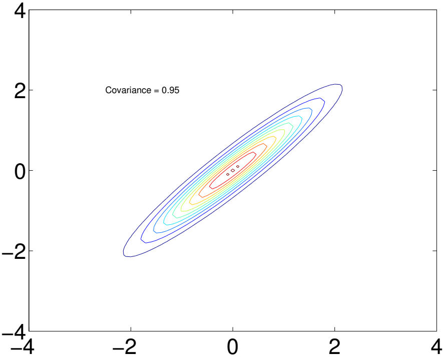
- If x_1 and x_2 are independent and have equal variance the contours of $p(x_1, x_2)$ form circles.
- However, if x_1 and x_2 are correlated the circles become stretched along the diagonal.



- How does this correlation effect the divergence (i.e. overlap) between a pair of distributions.

Correlation

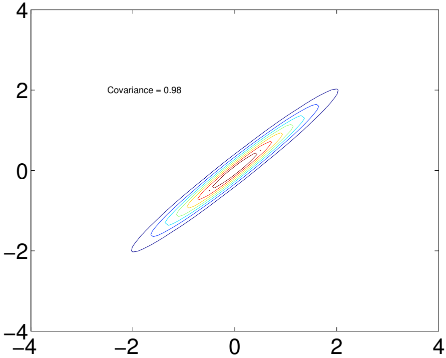
- If x_1 and x_2 are independent and have equal variance the contours of $p(x_1, x_2)$ form circles.
- However, if x_1 and x_2 are correlated the circles become stretched along the diagonal.



- How does this correlation effect the divergence (i.e. overlap) between a pair of distributions.

Correlation

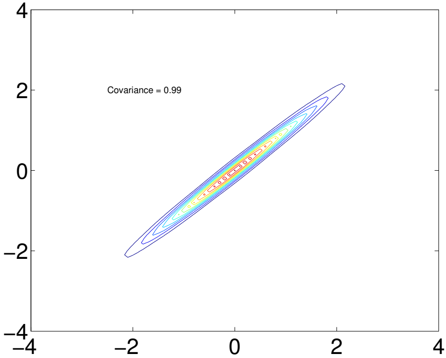
- If x_1 and x_2 are independent and have equal variance the contours of $p(x_1, x_2)$ form circles.
- However, if x_1 and x_2 are correlated the circles become stretched along the diagonal.



- How does this correlation effect the divergence (i.e. overlap) between a pair of distributions.

Correlation

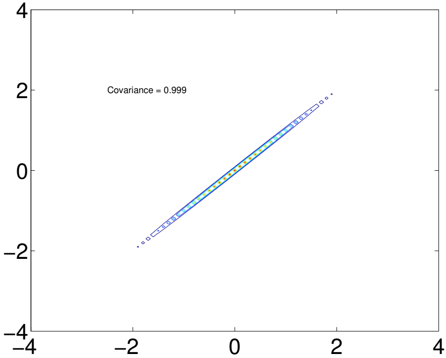
- If x_1 and x_2 are independent and have equal variance the contours of $p(x_1, x_2)$ form circles.
- However, if x_1 and x_2 are correlated the circles become stretched along the diagonal.



- How does this correlation effect the divergence (i.e. overlap) between a pair of distributions.

Correlation

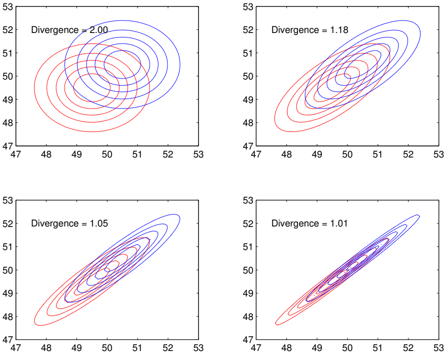
- If x_1 and x_2 are independent and have equal variance the contours of $p(x_1, x_2)$ form circles.
- However, if x_1 and x_2 are correlated the circles become stretched along the diagonal.



- How does this correlation effect the divergence (i.e. overlap) between a pair of distributions.

Divergence and correlation

- Class 1 - mean of x_1 = 49.5; mean of x_2 = 49.5
- Class 2 - mean of x_1 = 50.5; mean of x_2 = 50.5



- Distributions increasingly overlap as correlation between x_1 and x_2 increases, i.e. divergence decreases.
- So the divergence of the joint $p(x_1, x_2)$ distributions, cannot be predicted from divergence of distributions $p(x_1)$ and and divergence of distributions $p(x_2)$

Divergence for n-dimensional Gaussians

Remember,

- 1-d Gaussians are defined by their mean, μ, and a their variance σ^2.
- n-d Gaussians are defined by a vector of means, $\boldsymbol{\mu}$, and a covariance matrix, Σ.

The divergence between a pair of n-d Gaussians can be shown to be,

$$d_{12} = \frac{1}{2} trace\left(\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1 - 2I\right) + \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\Sigma_1^{-1} + \Sigma_2^{-1})(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

- $trace(X)$ is the sum of the values on the leading diagonal of matrix X
- $\mathbf{x}^T$ means the transpose of vector $\mathbf{x}$, i.e. turns a row vector into a column and vice versa.

1-d versus n-d Gaussian Divergence

Compare the equation for n-d Gaussian divergence,

$$d_{12} = \frac{1}{2} trace \left(\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1 - 2I\right) + \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\Sigma_1^{-1} + \Sigma_2^{-1})(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

With equation for the 1-d Gaussian divergence.

$$d_{12} = \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2} - 2\right) + \frac{1}{2}(\mu_1 - \mu_2)^2\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)$$

The 1-d equation is just a special case of the N-d equation.

1-d versus n-d Gaussian Divergence

The divergence between a pair of n-d Gaussians is,

$$d_{12} = \frac{1}{2} trace \left(\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1 - 2I \right) + \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T (\Sigma_1^{-1} + \Sigma_2^{-1})(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

Note, **if the features are uncorrelated**, (i.e. Σ are diagonal matrices), then the n-d divergence reduces to the sum of the 1-d divergences.

$$d_{12} = \sum_i \left(\frac{1}{2} \left(\frac{\sigma_{i1}^2}{\sigma_{i2}^2} + \frac{\sigma_{i2}^2}{\sigma_{i1}^2} - 2 \right) + \frac{1}{2}(\mu_{i1} - \mu_{i2})^2 \left(\frac{1}{\sigma_{i1}^2} + \frac{1}{\sigma_{i2}^2} \right) \right)$$

Selecting features, reconsidered

- Returning to our example and considering again picking the best **pair** of features,

left arm length (LAL)	2.5
right arm length (RAL)	2.4
hair length (HL)	1.8
weight (W)	1.3
shoe size (SS)	1.2
eye colour (EC)	0.01[2]

- LAL and RAL are clearly going to be highly correlated – people are pretty symmetrical!
- So it is possible that the joint distribution of LAL and HL would produce a higher divergence.

[2] These figures are made up for sake of illustration and are not realistic!

Selecting features, reconsidered

- Returning to our example and considering again picking the best **pair** of features,

left arm length (LAL)	2.5
right arm length (RAL)	2.4
hair length (HL)	1.8
weight (W)	1.3
shoe size (SS)	1.2
eye colour (EC)	0.01[2]

- LAL and RAL are clearly going to be highly correlated – people are pretty symmetrical!
- So it is possible that the joint distribution of LAL and HL would produce a higher divergence.

[2]These figures are made up for sake of illustration and are not realistic!

Selecting features, reconsidered

- Returning to our example and considering again picking the best **pair** of features,

left arm length (LAL)	2.5
right arm length (RAL)	2.4
hair length (HL)	1.8
weight (W)	1.3
shoe size (SS)	1.2
eye colour (EC)	0.01[2]

- LAL and RAL are clearly going to be highly correlated – people are pretty symmetrical!
- So it is possible that the joint distribution of LAL and HL would produce a higher divergence.

[2] These figures are made up for sake of illustration and are not realistic!

Simplest feature selection

- Simple scalar feature selection
  - Choose a 1-dimensional class separability criteria, C. e.g. could choose divergence, but there are others.
  - The value of the criterion $C(k)$ is computed for each feature, k.
  - Select the n features corresponding to the n best values of $C(k)$.
- Simple to perform but does not consider correlation between features.

Simplest feature selection

- Simple scalar feature selection
  - Choose a 1-dimensional class separability criteria, C. e.g. could choose divergence, but there are others.
  - The value of the criterion $C(k)$ is computed for each feature, k.
  - Select the n features corresponding to the n best values of $C(k)$.
- Simple to perform but does not consider correlation between features.

Simplest feature selection

- Simple scalar feature selection
  - Choose a 1-dimensional class separability criteria, C. e.g. could choose divergence, but there are others.
  - The value of the criterion $C(k)$ is computed for each feature, k.
  - Select the n features corresponding to the n best values of $C(k)$.
- Simple to perform but does not consider correlation between features.

Simplest feature selection

- Simple scalar feature selection
  - Choose a 1-dimensional class separability criteria, C. e.g. could choose divergence, but there are others.
  - The value of the criterion $C(k)$ is computed for each feature, k.
  - Select the n features corresponding to the n best values of $C(k)$.
- Simple to perform but does not consider correlation between features.

Simplest feature selection

- Simple scalar feature selection
  - Choose a 1-dimensional class separability criteria, C. e.g. could choose divergence, but there are others.
  - The value of the criterion $C(k)$ is computed for each feature, k.
  - Select the n features corresponding to the n best values of $C(k)$.
- Simple to perform but does not consider correlation between features.

Improved scalar feature selection algorithm

- As before, choose first feature as one with highest $C(k)$. Say it has index i_1
- Compute correlation, $\rho_{i_1 j}$ between selected feature, i_1, and each remaining feature, j.
- Compute **adjusted** separability between selected features and remaining features, according to,

$$C'(j) = \alpha_1 C(j) - \alpha_2 |\rho_{ij}| \qquad (3)$$

  where α_1 and α_2 are weighting factors giving relative importance of 1-d separability and correlation.
- Select 2nd feature as that which has highest $C'(j)$.
- Select nth feature as that which has highest,

$$\alpha_1 C(j) - \frac{\alpha_2}{n-1} \sum_{r=1}^{n-1} |\rho_{i_r j}| \qquad (4)$$

Improved scalar feature selection algorithm

- As before, choose first feature as one with highest $C(k)$. Say it has index i_1
- Compute correlation, $\rho_{i_1 j}$ between selected feature, i_1, and each remaining feature, j.
- Compute **adjusted** separability between selected features and remaining features, according to,

$$C'(j) = \alpha_1 C(j) - \alpha_2 |\rho_{ij}| \tag{3}$$

  where α_1 and α_2 are weighting factors giving relative importance of 1-d separability and correlation.
- Select 2nd feature as that which has highest $C'(j)$.
- Select nth feature as that which has highest,

$$\alpha_1 C(j) - \frac{\alpha_2}{n-1} \sum_{r=1}^{n-1} |\rho_{i_r j}| \tag{4}$$

Improved scalar feature selection algorithm

- As before, choose first feature as one with highest $C(k)$. Say it has index i_1
- Compute correlation, $\rho_{i_1 j}$ between selected feature, i_1, and each remaining feature, j.
- Compute **adjusted** separability between selected features and remaining features, according to,

$$C'(j) = \alpha_1 C(j) - \alpha_2 |\rho_{ij}| \quad (3)$$

  where α_1 and α_2 are weighting factors giving relative importance of 1-d separability and correlation.
- Select 2nd feature as that which has highest $C'(j)$.
- Select nth feature as that which has highest,

$$\alpha_1 C(j) - \frac{\alpha_2}{n-1} \sum_{r=1}^{n-1} |\rho_{i_r j}| \quad (4)$$

Improved scalar feature selection algorithm

- As before, choose first feature as one with highest $C(k)$. Say it has index i_1
- Compute correlation, $\rho_{i_1 j}$ between selected feature, i_1, and each remaining feature, j.
- Compute **adjusted** separability between selected features and remaining features, according to,

$$C'(j) = \alpha_1 C(j) - \alpha_2 |\rho_{ij}| \tag{3}$$

  where α_1 and α_2 are weighting factors giving relative importance of 1-d separability and correlation.
- Select 2nd feature as that which has highest $C'(j)$.
- Select nth feature as that which has highest,

$$\alpha_1 C(j) - \frac{\alpha_2}{n-1} \sum_{r=1}^{n-1} |\rho_{i_r j}| \tag{4}$$

Improved scalar feature selection algorithm

- As before, choose first feature as one with highest $C(k)$. Say it has index i_1
- Compute correlation, $\rho_{i_1 j}$ between selected feature, i_1, and each remaining feature, j.
- Compute **adjusted** separability between selected features and remaining features, according to,

$$C'(j) = \alpha_1 C(j) - \alpha_2 |\rho_{ij}| \tag{3}$$

  where α_1 and α_2 are weighting factors giving relative importance of 1-d separability and correlation.
- Select 2nd feature as that which has highest $C'(j)$.
- Select nth feature as that which has highest,

$$\alpha_1 C(j) - \frac{\alpha_2}{n-1} \sum_{r=1}^{n-1} |\rho_{i_r j}| \tag{4}$$

Feature vector selection

- Scalar techniques simple but have poor theoretical justification and will in general be sub-optimal.
- n-dimensional separability is poorly modeled by 1-d separability and pairwise correlations.
- We really want to consider the separability of the n-dimensional joint distributions directly.
- Consider a brute force approach which computes separability of joint distribution for all feature subsets.
- How many way of choosing n features from a possible m?

$$\frac{m!}{n!(m-n)!}$$

- This can be a very large number!

Feature vector selection

- Scalar techniques simple but have poor theoretical justification and will in general be sub-optimal.
- n-dimensional separability is poorly modeled by 1-d separability and pairwise correlations.
- We really want to consider the separability of the n-dimensional joint distributions directly.
- Consider a brute force approach which computes separability of joint distribution for all feature subsets.
- How many way of choosing n features from a possible m?

$$\frac{m!}{n!(m-n)!}$$

- This can be a very large number!

Feature vector selection

- Scalar techniques simple but have poor theoretical justification and will in general be sub-optimal.
- n-dimensional separability is poorly modeled by 1-d separability and pairwise correlations.
- We really want to consider the separability of the n-dimensional joint distributions directly.
- Consider a brute force approach which computes separability of joint distribution for all feature subsets.
- How many way of choosing n features from a possible m?

$$\frac{m!}{n!(m-n)!}$$

- This can be a very large number!

Feature vector selection

- Scalar techniques simple but have poor theoretical justification and will in general be sub-optimal.
- n-dimensional separability is poorly modeled by 1-d separability and pairwise correlations.
- We really want to consider the separability of the n-dimensional joint distributions directly.
- Consider a brute force approach which computes separability of joint distribution for all feature subsets.
- How many way of choosing n features from a possible m?

$$\frac{m!}{n!(m-n)!}$$

- This can be a very large number!

Feature vector selection

- Scalar techniques simple but have poor theoretical justification and will in general be sub-optimal.
- n-dimensional separability is poorly modeled by 1-d separability and pairwise correlations.
- We really want to consider the separability of the n-dimensional joint distributions directly.
- Consider a brute force approach which computes separability of joint distribution for all feature subsets.
- How many way of choosing n features from a possible m?

$$\frac{m!}{n!(m-n)!}$$

- This can be a very large number!

Feature vector selection

- Scalar techniques simple but have poor theoretical justification and will in general be sub-optimal.
- n-dimensional separability is poorly modeled by 1-d separability and pairwise correlations.
- We really want to consider the separability of the n-dimensional joint distributions directly.
- Consider a brute force approach which computes separability of joint distribution for all feature subsets.
- How many way of choosing n features from a possible m?

$$\frac{m!}{n!(m-n)!}$$

- This can be a very large number!

Sequential backward selection

- Finds good feature selection without trying all possibilities: start with all features and progressively remove one feature at a time.
- Consider an example where we have $m = 4$ features x_1, x_2, x_3, x_4.
  - Pick a class separability measure, C, and compute its value for the feature vector $[x_1, x_2, x_3, x_4]$
  - Eliminate one feature at a time and compute C for each possible 3-d vector, i.e. $[x_2, x_3, x_4]$, $[x_1, x_3, x_4]$, $[x_1, x_2, x_4]$ and $[x_1, x_2, x_3]$. Select the best, say, $[x_1, x_3, x_4]$.
  - Eliminate another feature and recompute C for each 2-d vector, i.e. $[x_3, x_4]$, $[x_1, x_4]$, $[x_1, x_3]$. Select the best…etc
- Note, this is a *suboptimal* approach. It does not guarantee to find the best selection/

Sequential backward selection

- Finds good feature selection without trying all possibilities: start with all features and progressively remove one feature at a time.
- Consider an example where we have $m = 4$ features x_1, x_2, x_3, x_4.
  - Pick a class separability measure, C, and compute its value for the feature vector $[x_1, x_2, x_3, x_4]$
  - Eliminate one feature at a time and compute C for each possible 3-d vector, i.e. $[x_2, x_3, x_4]$, $[x_1, x_3, x_4]$, $[x_1, x_2, x_4]$ and $[x_1, x_2, x_3]$. Select the best, say, $[x_1, x_3, x_4]$.
  - Eliminate another feature and recompute C for each 2-d vector, i.e. $[x_3, x_4]$, $[x_1, x_4]$, $[x_1, x_3]$. Select the best. . . etc
- Note, this is a *suboptimal* approach. It does not guarantee to find the best selection/

Sequential backward selection

- Finds good feature selection without trying all possibilities: start with all features and progressively remove one feature at a time.
- Consider an example where we have $m = 4$ features x_1, x_2, x_3, x_4.
  - Pick a class separability measure, C, and compute its value for the feature vector $[x_1, x_2, x_3, x_4]$
  - Eliminate one feature at a time and compute C for each possible 3-d vector, i.e. $[x_2, x_3, x_4]$, $[x_1, x_3, x_4]$, $[x_1, x_2, x_4]$ and $[x_1, x_2, x_3]$. Select the best, say, $[x_1, x_3, x_4]$.
  - Eliminate another feature and recompute C for each 2-d vector, i.e. $[x_3, x_4]$, $[x_1, x_4]$, $[x_1, x_3]$. Select the best. . . etc
- Note, this is a *suboptimal* approach. It does not guarantee to find the best selection/

Sequential backward selection

- Finds good feature selection without trying all possibilities: start with all features and progressively remove one feature at a time.
- Consider an example where we have $m = 4$ features x_1, x_2, x_3, x_4.
  - Pick a class separability measure, C, and compute its value for the feature vector $[x_1, x_2, x_3, x_4]$
  - Eliminate one feature at a time and compute C for each possible 3-d vector, i.e. $[x_2, x_3, x_4]$, $[x_1, x_3, x_4]$, $[x_1, x_2, x_4]$ and $[x_1, x_2, x_3]$. Select the best, say, $[x_1, x_3, x_4]$.
  - Eliminate another feature and recompute C for each 2-d vector, i.e. $[x_3, x_4]$, $[x_1, x_4]$, $[x_1, x_3]$. Select the best. . . etc
- Note, this is a *suboptimal* approach. It does not guarantee to find the best selection/

Sequential backward selection

- Finds good feature selection without trying all possibilities: start with all features and progressively remove one feature at a time.
- Consider an example where we have $m = 4$ features x_1, x_2, x_3, x_4.
  - Pick a class separability measure, C, and compute its value for the feature vector $[x_1, x_2, x_3, x_4]$
  - Eliminate one feature at a time and compute C for each possible 3-d vector, i.e. $[x_2, x_3, x_4]$, $[x_1, x_3, x_4]$, $[x_1, x_2, x_4]$ and $[x_1, x_2, x_3]$. Select the best, say, $[x_1, x_3, x_4]$.
  - Eliminate another feature and recompute C for each 2-d vector, i.e. $[x_3, x_4]$, $[x_1, x_4]$, $[x_1, x_3]$. Select the best. . . etc
- Note, this is a *suboptimal* approach. It does not guarantee to find the best selection/

Sequential backward selection

- Finds good feature selection without trying all possibilities: start with all features and progressively remove one feature at a time.
- Consider an example where we have $m = 4$ features x_1, x_2, x_3, x_4.
  - Pick a class separability measure, C, and compute its value for the feature vector $[x_1, x_2, x_3, x_4]$
  - Eliminate one feature at a time and compute C for each possible 3-d vector, i.e. $[x_2, x_3, x_4]$, $[x_1, x_3, x_4]$, $[x_1, x_2, x_4]$ and $[x_1, x_2, x_3]$. Select the best, say, $[x_1, x_3, x_4]$.
  - Eliminate another feature and recompute C for each 2-d vector, i.e. $[x_3, x_4]$, $[x_1, x_4]$, $[x_1, x_3]$. Select the best. . . etc
- Note, this is a *suboptimal* approach. It does not guarantee to find the best selection/

Sequential forward selection

- Similar to sequential backward selection, but we start with a single feature and progressively add features.
- Consider an example where we have $m = 4$ features x_1, x_2, x_3, x_4.
  - Pick a class separability measure, C, and compute its value for each individual feature. Select the best. Say, x_2
  - Now computer C for all possible 2-d vectors formed using x_2 and one other feature, $[x_1, x_2]$, $[x_2, x_3]$ and $[x_2, x_4]$. Select the best, say, $[x_2, x_3]$.
  - Add another feature and recompute C for the resulting 3-d vectors, i.e. $[x_1, x_2 x_3]$ and $[x_2, x_3, x_4]$ . . . etc
- Again, this is a *suboptimal* approach.
- Choosing forward or backward depends on whether the desired number of features n is closer to m (choose backward) or to 1 (choose forward).

Sequential forward selection

- Similar to sequential backward selection, but we start with a single feature and progressively add features.
- Consider an example where we have $m = 4$ features x_1, x_2, x_3, x_4.
  - Pick a class separability measure, C, and compute its value for each individual feature. Select the best. Say, x_2
  - Now computer C for all possible 2-d vectors formed using x_2 and one other feature, $[x_1, x_2]$, $[x_2, x_3]$ and $[x_2, x_4]$. Select the best, say, $[x_2, x_3]$.
  - Add another feature and recompute C for the resulting 3-d vectors, i.e. $[x_1, x_2 x_3]$ and $[x_2, x_3, x_4] \ldots$ etc
- Again, this is a *suboptimal* approach.
- Choosing forward or backward depends on whether the desired number of features n is closer to m (choose backward) or to 1 (choose forward).

Sequential forward selection

- Similar to sequential backward selection, but we start with a single feature and progressively add features.
- Consider an example where we have $m = 4$ features x_1, x_2, x_3, x_4.
  - Pick a class separability measure, C, and compute its value for each individual feature. Select the best. Say, x_2
  - Now computer C for all possible 2-d vectors formed using x_2 and one other feature, $[x_1, x_2]$, $[x_2, x_3]$ and $[x_2, x_4]$. Select the best, say, $[x_2, x_3]$.
  - Add another feature and recompute C for the resulting 3-d vectors, i.e. $[x_1, x_2 x_3]$ and $[x_2, x_3, x_4]$... etc
- Again, this is a *suboptimal* approach.
- Choosing forward or backward depends on whether the desired number of features n is closer to m (choose backward) or to 1 (choose forward).

Sequential forward selection

- Similar to sequential backward selection, but we start with a single feature and progressively add features.
- Consider an example where we have $m = 4$ features x_1, x_2, x_3, x_4.
  - Pick a class separability measure, C, and compute its value for each individual feature. Select the best. Say, x_2
  - Now computer C for all possible 2-d vectors formed using x_2 and one other feature, $[x_1, x_2]$, $[x_2, x_3]$ and $[x_2, x_4]$. Select the best, say, $[x_2, x_3]$.
  - Add another feature and recompute C for the resulting 3-d vectors, i.e. $[x_1, x_2 x_3]$ and $[x_2, x_3, x_4] \ldots$ etc
- Again, this is a *suboptimal* approach.
- Choosing forward or backward depends on whether the desired number of features n is closer to m (choose backward) or to 1 (choose forward).

Sequential forward selection

- Similar to sequential backward selection, but we start with a single feature and progressively add features.
- Consider an example where we have $m = 4$ features x_1, x_2, x_3, x_4.
  - Pick a class separability measure, C, and compute its value for each individual feature. Select the best. Say, x_2
  - Now computer C for all possible 2-d vectors formed using x_2 and one other feature, $[x_1, x_2]$, $[x_2, x_3]$ and $[x_2, x_4]$. Select the best, say, $[x_2, x_3]$.
  - Add another feature and recompute C for the resulting 3-d vectors, i.e. $[x_1, x_2 x_3]$ and $[x_2, x_3, x_4]$...etc
- Again, this is a *suboptimal* approach.
- Choosing forward or backward depends on whether the desired number of features n is closer to m (choose backward) or to 1 (choose forward).

Sequential forward selection

- Similar to sequential backward selection, but we start with a single feature and progressively add features.
- Consider an example where we have $m = 4$ features x_1, x_2, x_3, x_4.
  - Pick a class separability measure, C, and compute its value for each individual feature. Select the best. Say, x_2
  - Now computer C for all possible 2-d vectors formed using x_2 and one other feature, $[x_1, x_2]$, $[x_2, x_3]$ and $[x_2, x_4]$. Select the best, say, $[x_2, x_3]$.
  - Add another feature and recompute C for the resulting 3-d vectors, i.e. $[x_1, x_2 x_3]$ and $[x_2, x_3, x_4] \ldots$ etc
- Again, this is a *suboptimal* approach.
- Choosing forward or backward depends on whether the desired number of features n is closer to m (choose backward) or to 1 (choose forward).

Sequential forward selection

- Similar to sequential backward selection, but we start with a single feature and progressively add features.
- Consider an example where we have $m = 4$ features x_1, x_2, x_3, x_4.
  - Pick a class separability measure, C, and compute its value for each individual feature. Select the best. Say, x_2
  - Now computer C for all possible 2-d vectors formed using x_2 and one other feature, $[x_1, x_2]$, $[x_2, x_3]$ and $[x_2, x_4]$. Select the best, say, $[x_2, x_3]$.
  - Add another feature and recompute C for the resulting 3-d vectors, i.e. $[x_1, x_2 x_3]$ and $[x_2, x_3, x_4] \ldots$ etc
- Again, this is a *suboptimal* approach.
- Choosing forward or backward depends on whether the desired number of features n is closer to m (choose backward) or to 1 (choose forward).

Floating Search Methods

- Overcomes problem of *nesting effect*
  - backward selection: once a feature is discarded it can't be reconsidered.
  - forward selection: once a feature is added it can't be removed.
- Floating search methods substantially more complicated but offer better performance.
- Won't cover here but see T&K p. 286 if interested.

Summary

- Divergence between distributions increases as more dimensions are added.
- N-d divergence is sum of 1-d divergences, *but only if the dimensions are independent*.
- Divergence less than sum if the distributions are correlated.
- Optimal feature selection: try all subsets of features and pick the one the maximises class separability.
- For most problems too many features to consider this approach.
- A sequential feature selection algorithm can usually find a good set of features without having to consider all possible subsets.

Summary

- Divergence between distributions increases as more dimensions are added.
- N-d divergence is sum of 1-d divergences, *but only if the dimensions are independent*.
- Divergence less than sum if the distributions are correlated.
- Optimal feature selection: try all subsets of features and pick the one the maximises class separability.
- For most problems too many features to consider this approach.
- A sequential feature selection algorithm can usually find a good set of features without having to consider all possible subsets.

Summary

- Divergence between distributions increases as more dimensions are added.
- N-d divergence is sum of 1-d divergences, *but only if the dimensions are independent*.
- Divergence less than sum if the distributions are correlated.
- Optimal feature selection: try all subsets of features and pick the one the maximises class separability.
- For most problems too many features to consider this approach.
- A sequential feature selection algorithm can usually find a good set of features without having to consider all possible subsets.

Summary

- Divergence between distributions increases as more dimensions are added.
- N-d divergence is sum of 1-d divergences, *but only if the dimensions are independent*.
- Divergence less than sum if the distributions are correlated.
- Optimal feature selection: try all subsets of features and pick the one the maximises class separability.
- For most problems too many features to consider this approach.
- A sequential feature selection algorithm can usually find a good set of features without having to consider all possible subsets.

Summary

- Divergence between distributions increases as more dimensions are added.
- N-d divergence is sum of 1-d divergences, *but only if the dimensions are independent*.
- Divergence less than sum if the distributions are correlated.
- Optimal feature selection: try all subsets of features and pick the one the maximises class separability.
- For most problems too many features to consider this approach.
- A sequential feature selection algorithm can usually find a good set of features without having to consider all possible subsets.

Summary

- Divergence between distributions increases as more dimensions are added.
- N-d divergence is sum of 1-d divergences, *but only if the dimensions are independent*.
- Divergence less than sum if the distributions are correlated.
- Optimal feature selection: try all subsets of features and pick the one the maximises class separability.
- For most problems too many features to consider this approach.
- A sequential feature selection algorithm can usually find a good set of features without having to consider all possible subsets.