Lecture 15: Example exam questions

Nils Olovsson
nils.olofsson@igp.uu.se

Introduction to data science and Python programming for medical research
Uppsala University

2025

UPPSALA
UNIVERSITET

Content

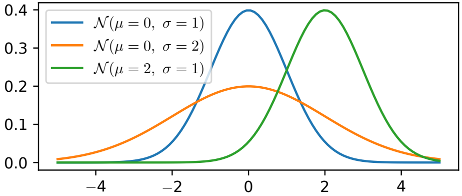
Questions

Questions

Questions

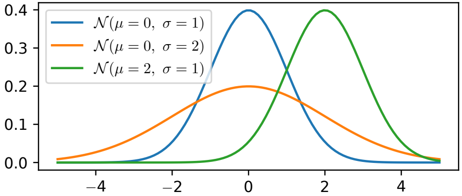
Questions: Summary statistics

What are the two parameters that define a normal distribution?



Questions: Summary statistics

What are the two parameters that define a normal distribution?

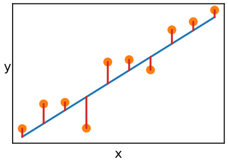


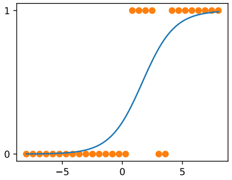
Answer:

Its center, or mean value, and the variance or standard deviation.

Questions: Regression

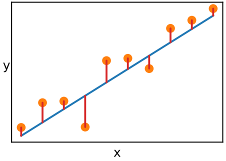
What is the difference in type of problem addressed between (linear) regression and logistic regression?

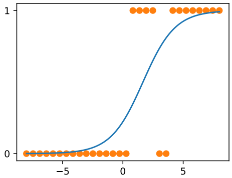




Questions: Regression

What is the difference in type of problem addressed between (linear) regression and logistic regression?





Answer:
Linear regression represents a model with continuous input and continuous output.
Logistic regression is used for classification.

Questions: Covariance matrix

A covariance matrix, C, always has certain properties. Use your knowledge of these to fill in the missing values in the following covariance matrix.

$$C = \begin{pmatrix} v_x & \square & b \\ a & v_y & c \\ \square & \square & v_z \end{pmatrix}$$

Questions: Covariance matrix

A covariance matrix, C, always has certain properties. Use your knowledge of these to fill in the missing values in the following covariance matrix.

$$C = \begin{pmatrix} v_x & \square & b \\ a & v_y & c \\ \square & \square & v_z \end{pmatrix}$$

Answer:

The covariance matrices are symmetric around the diagonal.

$$C = \begin{pmatrix} v_x & a & b \\ a & v_y & c \\ b & c & v_z \end{pmatrix}$$

Questions: Clustering

Kmeans and Gaussian mixture models (GMM) have several properties in common.
Mention <u>one</u> thing that is <u>similar</u> for these two methods and <u>one</u> thing that is <u>different</u> between them.

Questions: Clustering

Kmeans and Gaussian mixture models (GMM) have several properties in common. Mention one thing that is similar for these two methods and one thing that is different between them.

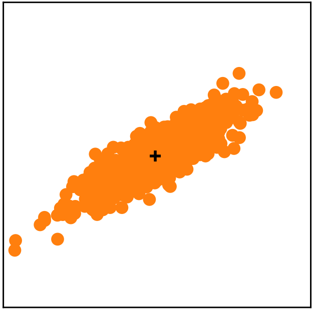
Answer:

1.
   - Both are limited to linear clusters.
   - Both are fit using expectation maximization (EM)
   - Both are parametrized by the centers of their clusters.
   - Both need the number of clusters as input setting.
2.
   - GMM is parametrized by the cluster centers and the covariance of a Gaussian function centered on each cluster.
   - GMM can cluster elongated data.

Questions: PCA

In the following 2D distribution the mean value is indicated.
Draw the two principal components of the data. Indicate which should be labeled the first, PC1, and second, PC2, principal component.

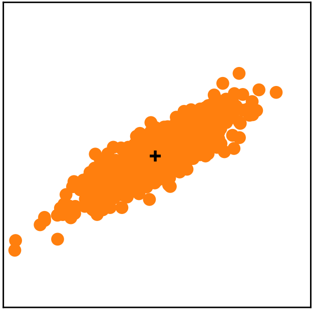
(The lengths of your drawn PCs to not have to correspond to anything in particular.)



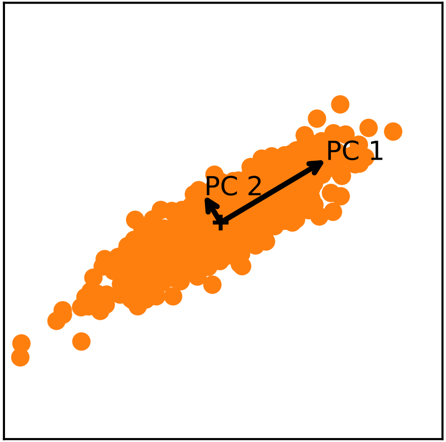
Questions: PCA

In the following 2D distribution the mean value is indicated.
Draw the two principal components of the data. Indicate which should be labeled the first, PC1, and second, PC2, principal component.

(The lengths of your drawn PCs to not have to correspond to anything in particular.)



Answer:



Questions: PCA variance

The following array contains the relative explained variances described by each principal component in a dataset of five features.

$$\begin{pmatrix} 0.50 & 0.30 & 0.11 & 0.05 & 0.04 \end{pmatrix}$$

How many components are needed in order to get a model that keeps at least 90% of the variance? Explain your reasoning.

Questions: PCA variance

The following array contains the relative explained variances described by each principal component in a dataset of five features.

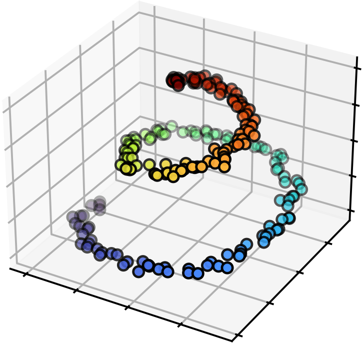
$$\begin{pmatrix} 0.50 & 0.30 & 0.11 & 0.05 & 0.04 \end{pmatrix}$$

How many components are needed in order to get a model that keeps at least 90% of the variance? Explain your reasoning.

Answer:
Three components are needed.
$0.50 + 0.30 + 0.11 = 0.91 = 91\%$ explained variance.

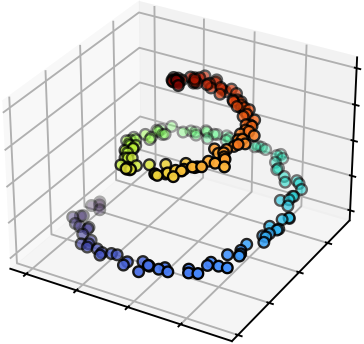
Questions: Manifold learning

By inspecting the data in the figure, what would you say is the latent number of dimensions of the data?
Explain your reasoning.



Questions: Manifold learning

By inspecting the data in the figure, what would you say is the latent number of dimensions of the data?
Explain your reasoning.



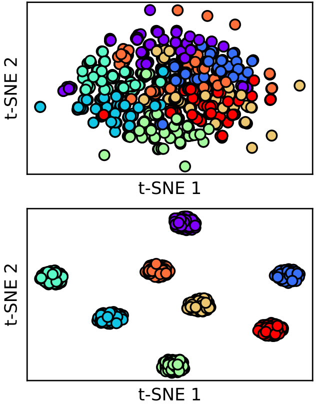
Answer:
1D. The data consists of a one dimensional *line* in the shape of a spiral in 3D.

Questions: t-SNE

Like most estimators/models in sklearn t-SNE has several settings specified as parameters when an instance is created.
Which parameter would you adjust to go from the first plot to the second?

TSNE

```
class sklearn.manifold.TSNE(n_components=2, *, perplexity=30.0,
early_exaggeration=12.0, learning_rate='auto', max_iter=None,
n_iter_without_progress=300, min_grad_norm=1e-07, metric='euclidean',
metric_params=None, init='pca', verbose=0, random_state=None,
method='barnes_hut', angle=0.5, n_jobs=None, n_iter='deprecated')
```

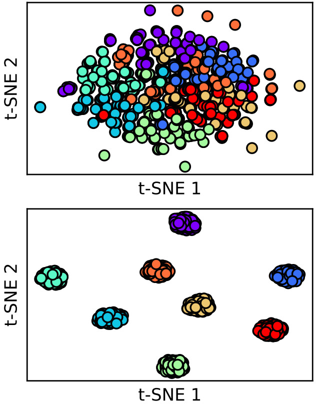


Questions: t-SNE

Like most estimators/models in sklearn t-SNE has several settings specified as parameters when an instance is created.
Which parameter would you adjust to go from the first plot to the second?

TSNE

```
class sklearn.manifold.TSNE(n_components=2, *, perplexity=30.0,
early_exaggeration=12.0, learning_rate='auto', max_iter=None,
n_iter_without_progress=300, min_grad_norm=1e-07, metric='euclidean',
metric_params=None, init='pca', verbose=0, random_state=None,
method='barnes_hut', angle=0.5, n_jobs=None, n_iter='deprecated')
```



Answer:
Perplexity.

Questions: SVM and random forests

For what type of data would SVM or random forests be a better choice than linear or logistic regression?

Questions: SVM and random forests

For what type of data would SVM or random forests be a better choice than linear or logistic regression?

Answer:
Nonlinear data.

Questions: Nonlinear models

Describe how you would identify if your model has been overfitted to the data.

Questions: Nonlinear models

Describe how you would identify if your model has been overfitted to the data.

Answer:

Split the data into a training and test set. Use a metric or score to measure the performance of the model on both the training and the test set to see if the model performs similarly on both.

Questions: Images (1)

What does image segmentation mean?

Questions: Images (1)

What does image segmentation mean?

Answer:
Annotation of regions or pixels in image that belong to and object, *e.g.* a cell or an organ.

Questions: Images (2)

Mention one image segmentation method.

Questions: Images (2)

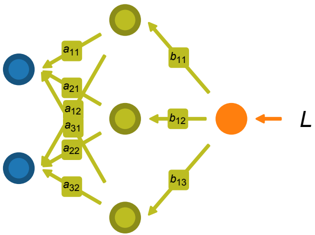
Mention one image segmentation method.

Answer:

- Threshold
- Otsu threshold
- Active contour
- Deep learning, convolutional neural networks.
- (Cluster methods)
- (Graph cuts)

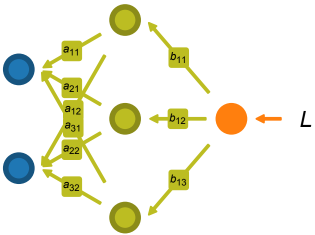
Questions: Neural network backpropagation

Neural networks are typically trained using backpropagation and some type of gradient descent based optimization.
What is it that is *backpropagated* through the network during training?



Questions: Neural network backpropagation

Neural networks are typically trained using backpropagation and some type of gradient descent based optimization.
What is it that is *backpropagated* through the network during training?



Answer:
The loss, error or discrepancy between the predicted output and the ground truth.

Questions: Neural network multiple layers

What is the purpose of having neural networks with more than a single layer?

Questions: Neural network multiple layers

What is the purpose of having neural networks with more than a single layer?

Answer:

It makes it possible to create nonlinear models using neural networks.

Questions: Deep learning image analysis

Mention one benefit of using convolutional layers instead of fully connected layers for image analysis.

Questions: Deep learning image analysis

Mention one benefit of using convolutional layers instead of fully connected layers for image analysis.

Answer:

Fully connected networks require a large number of connections/weights/parameters. Flattening images such that they can be used as input to a fully connected layer makes the image lose spatial connections between neighboring pixels.

Questions: Machine learning methods

Methods can be divided into supervised and unsupervised methods.

What is the difference between these two categories?

Name one method or family of methods for each of these two categories.

Questions: Machine learning methods

Methods can be divided into supervised and unsupervised methods.

What is the difference between these two categories?

Name one method or family of methods for each of these two categories.

Answer:

Supervised methods should *predict* some output given a data point as input.
Needs both a data matrix and target vector for model fitting.

<u>Supervised:</u> Linear regression, logistic regression, SVM, Random forests, neural networks.

<u>Unsupervised:</u> Clustering, PCA, Manifold learning.

Questions: Supervised vs unsupervised

Name a general use case for unsupervised learning methods.

Questions: Supervised vs unsupervised

Name a general use case for unsupervised learning methods.

Answer:
Dimensionality reduction, visualization, data exploration.