Topic 1 - Machine learning

Difference between supervised and unsupervised
-Supervised teach the computer to do something -presenting examples, have labels. Classifications, feature selection, regression, CNN -tumor example

-Unsupervised-do it by himself- no labels- large computer clusters, the computer finds the structure itself, no idea about the results, no feedback based on prediction. Clustering (GMM), dimensionality reduction, (PCA) kernel smoothing. example:cocktail party algorithm

-third type-reinforcement learning- no supervisor but reward. Feedback based on the reward. Exploration and exploitation and prediction vs control.

Overfitting- analysis that corresponds too close/exactly to a specific set of data and may therefore fail to fit the additional data. Low training error but high testing error. To avoid, we do cross validation, or regularization.

N-fold cross validation- data partitioned into N groups equally sized groups. You train on all but one group and test on the one not trained on. This is done for all groups. Cross validation estimate how the accurate predictive model will perform in practice, and gives mean error value.
1- Mean error value gives accuracy
Dividing each group into training set, validation set, and test set.
Higher the N number, higher the computational cost.

Accurancy: TP+TN/TP+FP+FN+TN
F-measure:The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst at 0. Often more useful than accuracy, especially if uneven class distribution. 2 x precision x recall/precison +recall
Recall:tp/tp+fn
Precision:tp/tp+fp

Topic 2- Basic Probability Theory

Joint probability indicates the probability of all combinations of 2+ variables, while **conditional probability** is that one event occurs if the other occurs too. They relate to each other, because when calculating the conditional probability, we need the joint probability divided by marginal valuer (example:what is the conditional probability that any male is cheating= cheating male/ total of cheating people (man +woman)

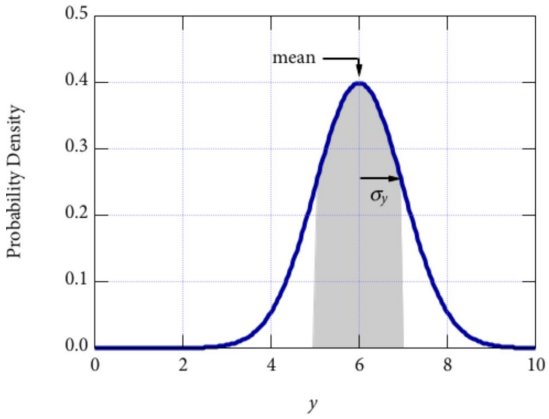
Bayer's Theorem or Bayes' rule describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)},$$

A;B=events
P(A|B)= conditional probability likehood of A occuring if B true
P(B|A)=same for B vice versa

P(A) + P(B) -probabilities of observing A and B independently from each other…

gaussian..



Variance is how wide it is.

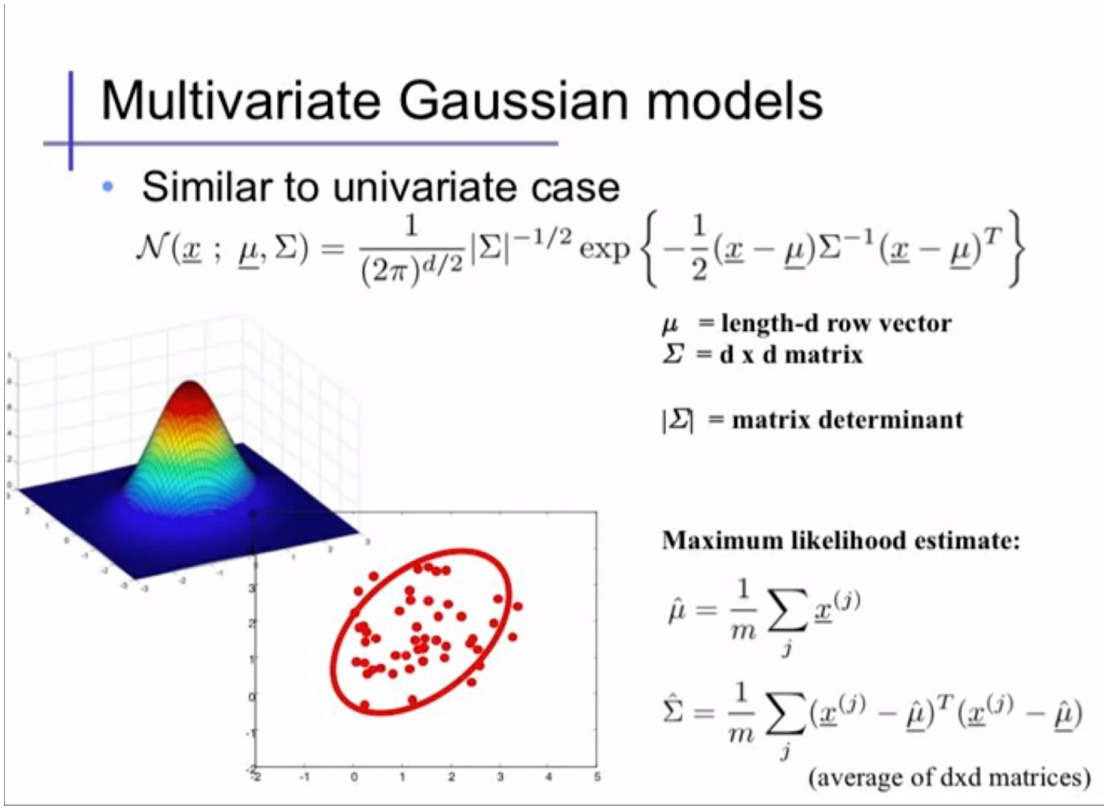
Generalization of the **variance for multivariate data**?
-Is Covariance. Matrix; how features are relating between each other.

Covariance matrix look: N-dimensional data.. n*n size, square matrix with diagonal values 1. Diagonally-symmetric. 0 means the two variables are independent.

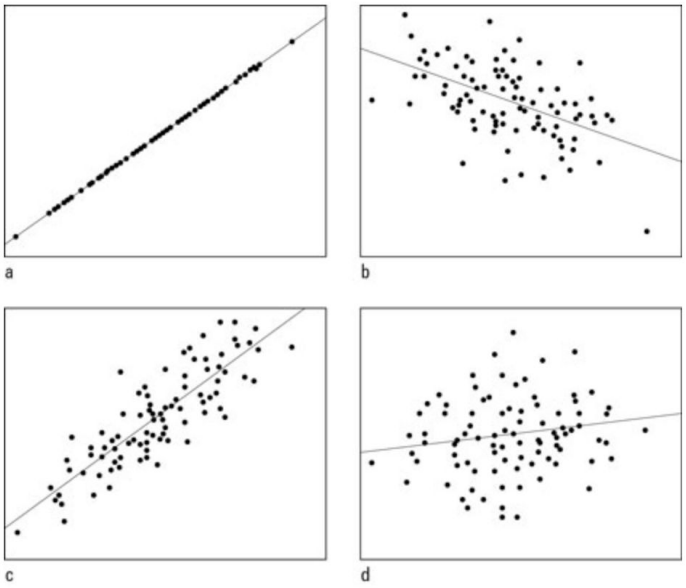
Mixture of the multiple Gaussians.. 2 or more gaussians together..**Gaussian Mixture Model** (GMM) is a parametric probability density function represented as a weighted sum of **Gaussian** component densities. ... GMM parameters are estimated from training data using the iterative Expectation-Maximization (EM) algorithm or Maximum A Posteriori (MAP) estimation from a well-trained prior **model**.

Univariate Gaussian is described by Mean,variance (distribution of only one random variate)
Multivariate is described by covariance (distribution of multiple variables or a group of them)

Equation of a multivariate Gaussian:



Topic 3 Correlation



Scatterplots with correlations of a) +1.00; b) –0.50; c) +0.85; and d) +0.15.

1. Linear correlation between two features is +1..0 or -1..

a)+1 in case of a perfect direct increasing
b)-1 perfect decrease Inverse(anti correlation)
c)0 - independent variables,convers not true because the correlation coefficient detects only linear dependencies between variables

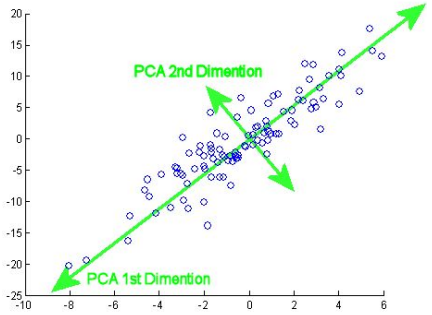
3. Range correlation value can have..
-1<=r<=1 .. 0-no correlation

4. Equation for the correlation..

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

Topic 4 PCA… best projection.. Best variance preservation.. Compress and visualize data

1. Indicate axes in which scale..
2. Angle between 2 eigenvectors..- 90 degrees..



3. lost/preserved projecting on n-best first, for ow much is responsible every eigenvector..
4. the cumulative percentage=eigenvalue/sum of eigenvalues
5. Eigenvalue decomposition is part of a PCA..We calculate it to find out which eigenvalues represent our dataset the most.

Topic 5.- Classification

MDC,LDA,QDA.. Difference..

**** the structure and number of different covariance matrices that determine the classifier. (*) QDA? LDA?***

MDC identity covariance matrix only
LDA mean and covariance matrix
QDA independent covariance matrix

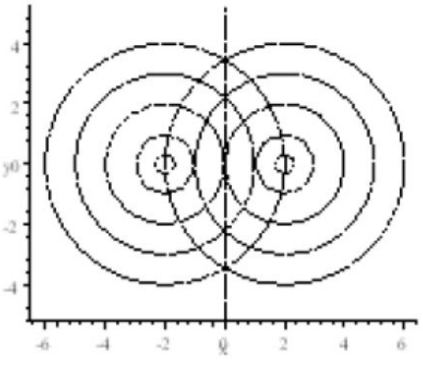
(MDC- [1 0 0 1] -Identity
LDA- cluster + cluster-mean cov. Matrix..
QDA- not dependant..)
Generally QDA assumes much, and LDA little

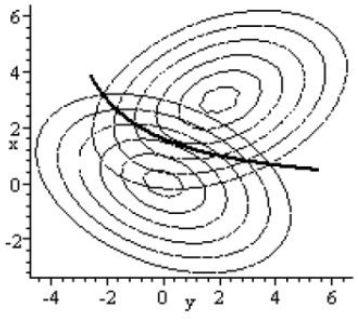
MDC - Data of each class are drawn from a radialsymmetric Gaussian distribution with the identical covariance matrix of the form Σ = σ 2 I.
LDA - Data of each class are drawn from a Gaussian distribution with identical covariance matrix Σ(sigma) of any form.
QDA - Data of each class k are drawn from a Gaussian distribution with its class-specific covariance matrices Σk(sigma k) of any form

2. Data points visualization

- ***Given data points of two classes, be able to visualize the estimated covariance matrices (by level lines) for MDC, LDA, or QDA. Explain your visualization. (*)***

I believe this is what he wants :

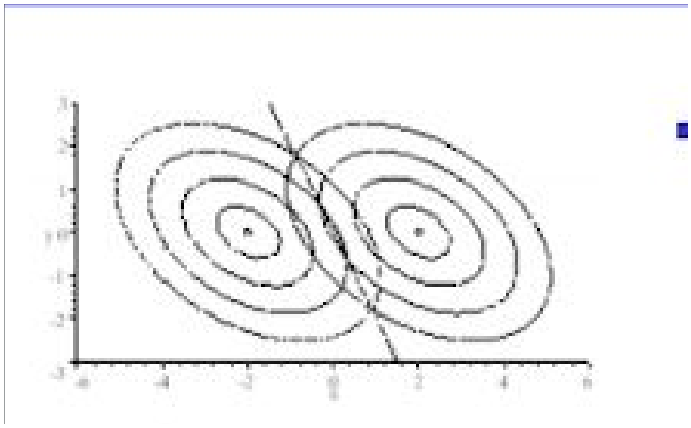




QDC

MDC

AAAAAANAAA nice, he likes to express himself very fancy :) :)) but can't find for linear, ***-<u>for linear is also straight line.</u>***



=minimum malhalanobis classifier

Bayesian classification-drawing.. Linearily separated classes. (1,0)Towards the higher probability distribution

Parametric classifiers :estimate parameters,,LDA,qda,mda... Simple,less data, speed..but constrained, limited complexity and poor fit
steps:

1. Select a form for the function.
2. Learn the coefficients for the function from the training data

Non parametric :do not make strong assumptions,free to learn any form.(knn, decision trees).-Flexibility: Capable of fitting a large number of functional forms.
Power: No assumptions (or weak assumptions) about the underlying function. Performance: Can result in higher performance models for prediction. But more data to process, slower.

.knn-graphic-non parametric.. Used for classification and regressions.. Classified objects by a majority vote of its neighbours.. If k=1 objects is assigned to the class of that single N..

Neural network layers:Fully connected layer- looks like an output from previous layer and determine next..(previous which features most correlate to particular class..the output from one neuron is the weighted sum of the values from the previous neurons
b)convolutional-first layer..
c)output. Last- our classifier.. Value for each class.. the output layer is often a Softmax layer from which the class with highest probability is chosen

convolution
Kernel-identity features, (example: dark pix value=1, light pix value =0)

Neural networks-how is the activity for a unit in a feature map calculated for a convolutional and for a fully connected layer?- A unit in a fully connected layers takes a weighted sum of every unit in the previous layer and applies it to an activation function (activation function in the current layer-I understand). For

convolutional layer, instead of being connected to every unit from the previous layer, a unit in the next layer is only connected to a "window" of units in the previous layer

size of a feature map depending on input,kernel size,padding and stride- In a 32×32 image , dragging the 5×5 receptive field across the input image data with a stride width of 1 will result in a feature map of 28×28 $(32-5+1 \times 32-5+1)$

softmax-,sigmoid function- nonlinear, output between 0/1. Function is used in final layer of CNN.

Dropout = in each iteration certain % of data is loss. **batch normalization** = helps with faster learning and speed, reduce internal covariance shift in neural networks

transfer learning in neural networks-existing pretrained network to solve a new problem..results in a faster training; Transfer learning is when you initialize your model with pretrained weights and then train your model on your own data.

How does neural network learn? Supervised learning-attempt to solve an opimalization problem to choose a self of parameters that minimize an error function..improve performance over time.. Bias, building knowledge…

Gradient descent-algorithm for finding local minimum of a function.. Takes steps proportional to the negative of the gradient of the function to the current rates of lower error..

delta rule**= Delta rule** is a gradient descent learning rule for updating the weights of the inputs to artificial neurons in a single-layer neural network.

difference between classification and regression.. Both are supervised learning.
Regression-predicts the **continuous** output value using training data..(univariate, linear, multivariate) examples: ebola spread, mortality rate, height,..
Classification-grouping output into a class, each class having a **discrete** variable[0 1 .. binary], categorical. Cancer or not, Skin pixel or not, Iris dataset,..

Topic 6 Clustering

K means works for organizing data into groups… analyze organically formed groups.. Calculating mean of each cluster, and each point reassigned to the mean with least distance. Repeated until convergence.

2.DETERMINE THE BEST NUMBER OF CLUSTERs-cross validation.. Subjective.. Depends on the shape.. Scale of the distribution of the points in a data… increase in “k’ without penalty will reduce the amount of error in the resulting clustering.. Elbow method-looks for % of variance..

3.. Variance ratio criterion of Czalinsky Harabasz tries to find the optimal number of cluster! (answer to previous Q?)

Topic 7-• Feature Selection (compared to PCA, it is supervised learning)

(=selecting a subset of relevant features (variables, predictors) for use in model construction)
– What is the difference between a wrapper and a filter? (*)

Filter Methods: the subset selection procedure is independent of the learning algorithm and is generally a pre-processing step. Faster learning pipeline.
-uses correlation between features and labels

Wrapper Methods: the subset selection takes place based on the learning algorithm used to train the model itself.
-uses classifier (Best accuracy could indicate best feature subset.)

– Explain forward and backward selection. (**)

Forward selection: data drive. model building approach; faster than searching over all subsets. Pipeline:

- Check every single of the 6 features in isolation.
- Evaluate it according to godness of subset criterion (more on that later).
- Select the best feature (according to subset criterion).
- Check each of the remaining 5 features and evaluate it together with the previously selected ones.
- Select the pair of features with best godness of subset criterion.
- Check each of the remaining 5 features together with the 2 already selected ones . . . until you check the entire set of 6 features.

Backward selection works through elimination instead.

TOPIC 8

• Regression

– Explain what linear regression is. (*)

Linear regression is a “linear” approach for estimating predictions for continuous variables, based on the already provided data. Examples of uses: forecasting, predictions (Ebola spread, consumption spending), error reductions(financial risks),... The relationship (between the data and the estimated data) is modeled using linear prediction function. Those models are made based on the least square approach.

-very sensitive to outliers; so we either filter them or use a “robust regression method”

– What are Least Mean Squares? (*)

Minimizing the sum of square residuals =difference between an observed value, and the fitted value provided by a model (mean)

$$\min_w \sqrt{\frac{1}{I}\sum_{i=1}^{I}(f(x_i; w) - y_i)^2}$$

– What is the Least Mean Squares algorithm? (***)

Using gradient descent algorithm

-starts with random “guess” number and apply least mean square

-repeat “update step” until convergence

Topic 9

• Reinforcement Learning (RL)

– What is RL? (*)

No supervisor, only a reward signal (Rt scalar feedback signal, indicates how well the agent is doing at step t. Agent job is to maximise the reward

Delayed feedback

Sequential

examples:Motor controls, robotics, games, business operations,

– What is Q-learning? (***)

Q-learning is a model-free reinforcement **learning** technique. Specifically, **Q-learning** can be used to find an optimal action-selection policy.

– What is Deep Q-learning? (***)

Deep Q learning is using a deep neural network to estimate the Q-function.