Course 495: Advanced Statistical Machine Learning/Pattern Recognition

• Lectures: Stefanos Zafeiriou

• Goal (Lectures): To present modern statistical machine learning/pattern recognition algorithms. The course focuses on statistical latent variable models (continuous & discrete).

• Goal (Tutorials): To provide the students the necessary mathematical tools for deeply understanding the models.

• Main Material: Pattern Recognition & Machine Learning by C. Bishop Chapters 1,2,12,13,8,9

• More materials in the website:
  http://ibug.doc.ic.ac.uk/courses/advanced-statistical-machine-learning-495/

• Email of the course: course495imperial@gmail.com
Statistical Machine Learning

The two main concepts are: **machine learning** and **statistics**

**Machine Learning**: A branch of Artificial Intelligence (A.I.) that focuses on designing, developing and studying the properties of algorithms that learn from data.

**Statistics**: A branch of applied mathematics that study collection, organization, analysis, interpretation, presentation and visualization of data and modelling their **randomness** and **uncertainty** using probability theory, as well as linear algebra and analysis.
Learn:

- Learning is at the core of the problem of Intelligence. Models of learning are used for understanding the function of the brain. Learning is used to develop modern intelligent machines.
- In many disciplines learning is now one of the main lines of research, i.e. signal/speech/ (medical) image processing, computer vision, control/robotics, natural language processing, bioinformatics etc.
- It starts to dominate other domains such as software engineering, security sciences, finance etc.
- Major players invest huge amount of money in modern learning systems (e.g., Deep Learning by Google and Facebook)
- Next 25 years will be the age of machine learning
Machine Learning in movies

Terminator 2
Machine Learning in movies

**Movie**: 2001: A Space Odyssey

**Director**: Stanley Kubrick

Hal 9000
Statistical Machine Learning

Some applications of modern learning algorithms

- Face detection
- Object/Face tracking
- Biometrics
- Speech/Gesture recognition
- Image segmentation
- Finance
- Bioinformatics
Applications

Object & face detection (modern cameras)
Applications

Face/Iris/Fingerprint recognition (biometrics)
Applications

Object-target tracking
Applications

Speech Recognition (voice Google search)

Waveform

Hello world
Applications

Gesture recognition (Kinect games)

Gestures
Applications

Image Segmentation
Applications

Biological data

Discover genes associated to a disease

what are you thinking?
What does the machine learn in each application?

Face Recognition: learn a classifier, i.e. a function (design of classifiers is covered in 424: Neural Computation)

But images are very high dimensional objects!!! We need to reduce them!!!
Latent Variable Models

Model (Deterministic):
\[ y = W^T x \]
Parameters: \( \theta = \{W\} \)

Model (probabilistic):
\[ x = W y + \mu + e \]
\[ e \sim N(e|0, \sigma^2 I) \]
\[ y \sim N(y|0, I) \]
Parameters: \( \theta = \{W, \mu, \sigma^2\} \)

Graphical Model:
\[ \theta \rightarrow x_n \]
\[ \{v_n\} \rightarrow y \]
Latent Variable Models

• Deterministic model:
  ✓ There is no randomness (uncertainty) associated with the model.
  ✓ We can compute the actual values of the latent space.

• Probabilistic model:
  ✓ We assign probability distributions to the latent variables and model their dependencies.
  ✓ We can compute only statistics of the latent variables.
  ✓ More flexible than deterministic models.

• Generative Probabilistic models:
  ✓ Model observations drawn from a probability density function (pdf) (i.e., model the way data are “generated”)

Stefanos Zafeiriou  Adv. Statistical Machine Learning (course 495)
Latent Variable Models

• Generative Probabilistic models:
  ✓ Model the complete (joint) likelihood of both the data and latent structure

• But what is the latent (hidden) structure:
  Intuitive.

Latent (hidden) structure
Latent Variable Models

Speech:

Word: need

Phonemes: n iy d

Image:

Object
Background

Latent structure
Latent Variable Models (Static)

General Concept:

\[
\begin{align*}
&x_1, x_2, x_3, \ldots, x_N \\
y_1, y_2, y_3, \ldots, y_N
\end{align*}
\]

Share a common linear structure

\[
x = Wy + \mu + e \\
e \sim N(e|0, \sigma^2 I) \\
y \sim N(y|0, I)
\]

We want to find the parameters:

\[
\theta = \{W, \mu, \sigma^2\}
\]

Joint likelihood maximization:

\[
p(x_1, \ldots, x_N, y_1, \ldots, y_N | \theta) = \prod_{i=1}^{N} p(x_i | y_i, W, \mu, \sigma) \prod_{i=1}^{N} p(y_i)
\]
Latent Variable Models (Dynamic, Continuous)
Latent Variable Models (Dynamic, Continuous)

Generative Model

\[ x_n = W y_n + e_n \]
\[ y_1 = \mu_0 + u \]
\[ y_n = A y_{n-1} + v_n \]

Parameters: \( \theta = \{ W, A, \mu_0, \Sigma, \Gamma, P_0 \} \)

Noise distribution

\[ e \sim N(e|0, \Sigma) \]
\[ u \sim N(u|0, P_0) \]
\[ v \sim N(v|0, \Gamma) \]
Latent Variable Models (Dynamic, Continuous)

Markov Property: \( p(\mathbf{y}_i, | \mathbf{y}_1, \ldots, \mathbf{y}_{i-1}) = p(\mathbf{y}_i | \mathbf{y}_{i-1}) \)

Joint likelihood:

\[
p(\mathbf{x}_1, \ldots, \mathbf{x}_T, \mathbf{y}_1, \ldots, \mathbf{y}_T) = \prod_{i=1}^{N} p(\mathbf{x}_i | \mathbf{y}_i, \mathbf{W}, \mathbf{\mu}_0, \Sigma) p(\mathbf{y}_1 | \mathbf{\mu}_0, \mathbf{P}_0) \prod_{i=2}^{N} p(\mathbf{y}_i | \mathbf{y}_{i-1}, \mathbf{A}, \Gamma)
\]
Latent Variable Models (Dynamic, Discrete)

Word: need

Phonemes: n iy d

Latent structure takes discrete values:
\( y_t \in \{\text{start}, n, iy, d, \text{end}\} \)
Latent Variable Models (Dynamic, Discrete)

\[
A = [a_{ij}] = [p(y_t|y_{t-1})]
\]

\[
\pi = [p(y_1)]
\]

\[
\begin{array}{c|cccccc}
\end{array}
\]
Latent Variable Models (Dynamic, Discrete)

Matrix of transition probabilities is represented as an automaton.

Data generation: e.g. if the latent variable is $\mathbf{y}_t = \mathbf{b}$, $\mathbf{x}_t \sim N(\mathbf{x}_t | \mathbf{m}_b, \mathbf{S}_b)$

Parameters $\theta = \{A, \{\mathbf{m}_b, \mathbf{S}_b\}_b, \pi\}$
Latent Variable Models (Dynamic, Discrete)

Word model

HMM of a word

Language model

HMM of a word

HMM of a word
Latent Variable Models (Spatial)

Image Segmentation

Brain tumour segmentation
Latent Variable Models (Spatial)

Undirected spatial dependencies

Markov Random fields
Latent Variable Models (Spatial)

Markov Random fields

\[ p(y_{11}, \ldots, y_{nm}) = \frac{1}{Z} \prod_c \psi(y_c) \]

\( C \) is the maximal clique.

Potential function: \( \psi(y_c) = e^{-E(y_c)} \)

Partition function: \( Z = \sum_y \prod_c \psi(y_c) \)

Markov blanket:

\[ p(y_{ul}, y_{vk} | Y / y_{ul}, y_{vk}) = p(y_{ul} | Y y_{ul}) p(y_{vk} | Y y_{vk}) \]

Complete likelihood:

\[ p(X, Y | \theta) = \frac{1}{Z} \prod_{uv} p(x_{uv} | y_{uv}, \theta) \prod_c \psi(y_c | \theta) \]
Summarize what we will study?

Deterministic Component Analysis (3 weeks)

Unsupervised approaches:
- Principal Component Analysis, Independent Component Analysis, Graph-based Component Analysis, Slow feature Analysis

Supervised approaches:
- Linear discriminant analysis

What we will learn?:
- How to find the latent space directly $\mathbf{y}$.
- How to find the latent space via linear projections $\mathbf{y} = \mathbf{W}^T \mathbf{x}$. 
Summarize what we will study?

Static data

Sequential data

Spatial data
Probabilistic Principal component analysis

What we will learn?:

• How to formulate probabilistically the problem.
• How to find both data moments $E[y_i]$, $E[yy^T]$ and parameters $\theta$
Summarize what we will study?

Sequential data (3 weeks):

What are the models?:
• *The Kalman filter (1 week)/ the particle filter (1 week).*
• *The Hidden Markov Model (1 week).*

What we will learn?:
• *How to formulate probabilistically the problems and learn parameters.*
Summarize what we will study?

Spatial data (2-3 weeks):

What are the models?:

- Gaussian Markov Random Fields (1 week).
- Discrete Markov Random Fields (1 week) and Mean Field Approximation.

What we will learn?:

- How to formulate probabilistically the problems and learn parameters.
What are the tools we need?

- We need elements from differential/integral calculus using vectors and matrices
  \[ \nabla_W f(W) = \frac{\partial f(W)}{dW_{ij}} \]

  ✓ *Matrix cookbook* (by Michael Syskind Pedersen & Kaare Brandt Petersen)

  ✓ Mike Brookes (EEE) has a nice page
    [http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/calculus.html](http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/calculus.html)

- We need linear algebra (matrix multiplications, matrix inversion etc). Special focus on eigenanalysis.

  ✓ Excellent book is *Matrix Computation* by Gene H. Golub, & Charles F. Van Loan
What are the tools we need?

- We need elements of optimization (mainly simple quadratic optimization problems with constraints which result to generalized eigenvalue problems). Refresh memory on how Lagrangian multipliers are used etc.

- We need tools from probability/statistics: random variable, probability density/mass function, marginalization
  - Assume pdf \( p(x) \) the probability is computed \( P(x \in A) = \int_A p(x)dx \)
  - Marginal distributions \( p(x) = \int_x p(x,y)dx \)
  - \( p(y) = \int_y p(x,y)dy \)
  - First and second order moments \( E(x) = \int_x xp(x)dx \)
  \( E(xx^T) = \int_x xx^T p(x)dx \)
What are the tools we need?

- Bayes rule and conditional independence.

\[ p(x, y) = p(x|y)p(y) \] Bayes rule

Conditional independence

\[ p(x, z, y) = p(x|y)p(z|y)p(y) \]

- Finally, we need tools from algorithms recursion and dynamic programming
What to have always in mind?

• What is my model?

• What are my model’s parameters? $\theta = \{W, \mu, \sigma^2\}$

• How do I find them? Maximum Likelihood

• How do I use the model?
Assignments

• Assessment (90% from written exams & 10% from assignments)

• Two assignments:

  One will be given next week and should be delivered by 21\textsuperscript{st} of February

  The second will be given 24\textsuperscript{th} of February and should be delivered 17\textsuperscript{th} of March
Adv. Statistical Machine Learning – Lectures

- Lecture 1-2: Introduction
- Lecture 3-4: A primer on calculus, linear algebra, probability/statistics
- Lecture 5-6: Deterministic Component Analysis (1)
- Lecture 7-8: Deterministic Component Analysis (2)
- Lecture 9-10: Deterministic Component Analysis (3)
- Lecture 11-12: Probabilistic Principal Component Analysis
- Lecture 13-14: Sequential Data: Kalman Filter (1)
- Lecture 15-16: Sequential Data: Kalman Filter (2)
- Lecture 17-18: Sequential Data: Hidden Markov Model (1)
- Lecture 18-19: Sequential Data: Hidden Markov Model (2)
- Lecture 19-20: Sequential Data: Particle Filtering (1)
- Lecture 20-21: Sequential Data: Particle Filtering (2)
- Lecture 22-23: Spatial Data: Gaussian Markov Random Field (GMRF)
- Lecture 23-24: Spatial Data: GMRF (2)
- Lecture 25-28: Spatial Data: Discrete MRF (Mean Field)
Machine Learning (models/parameters)

What does the machine learn in each application (parameters of a model)?

- Object detection: learn classifier, i.e. a function (design of classifiers is covered in Neural Computation)

Model: $y_i = w^T x_i + b$
Parameters: $\theta = \{w, b\}$