Machine Learning applied to neuroscience and cognition

Sophie Deneve
Angela Yu
# CA6b-Planning

Fridays 9h30: 12h30

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Presenter(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>13/02/2015</td>
<td>Bayesian approaches applied to cognition</td>
<td>Sophie Deneve</td>
</tr>
<tr>
<td>20/02/2015</td>
<td>Bayesian approaches applied to neural processing</td>
<td>Sophie Deneve</td>
</tr>
<tr>
<td>27/02/2015</td>
<td>TD on population coding</td>
<td>Charlotte Le Mouel</td>
</tr>
<tr>
<td>06/03/2015</td>
<td>Normative approaches applied to perception</td>
<td>Angela Yu</td>
</tr>
<tr>
<td>13/03/2015</td>
<td>TD on Bayesian methods</td>
<td>Izzet Burak Yildiz</td>
</tr>
<tr>
<td>20/03/2015</td>
<td>Supervised and unsupervised learning</td>
<td>Sophie Deneve</td>
</tr>
<tr>
<td>03/04/2015</td>
<td>Efficient coding in sensory systems</td>
<td>Sophie Deneve</td>
</tr>
<tr>
<td>10/04/2015</td>
<td>&gt; vacances</td>
<td></td>
</tr>
<tr>
<td>17/04/2015</td>
<td>TD on Sparse coding</td>
<td>David Chulz</td>
</tr>
<tr>
<td>24/04/2015</td>
<td>Introduction to graphical models</td>
<td>Sophie Deneve</td>
</tr>
<tr>
<td>01/05/2015</td>
<td>&gt; férié</td>
<td></td>
</tr>
<tr>
<td>08/05/2015</td>
<td>&gt; férié</td>
<td></td>
</tr>
<tr>
<td>15/05/2015</td>
<td>TD on belief propagation</td>
<td>Pantelis Leptougos</td>
</tr>
<tr>
<td>22/05/2015</td>
<td>Coding in balanced spiking networks</td>
<td>Sophie Deneve</td>
</tr>
<tr>
<td>29/05/2015</td>
<td>TD on balanced spiking networks</td>
<td>Erwan Ledoux</td>
</tr>
<tr>
<td>05/06/2015</td>
<td>Project presentation</td>
<td></td>
</tr>
</tbody>
</table>
Machine **learning** and artificial intelligence

“Classic” AI: Turing test

Verbal communication = ultimate intelligence criteria?
Machine learning and artificial intelligence

What is hard about playing chess?
Curse of Dimensionality

\[ \begin{align*}
D = 1 & \quad \text{(1 dimension)} \\
D = 2 & \quad \text{(2 dimensions)} \\
D = 3 & \quad \text{(3 dimensions)}
\end{align*} \]
Curse of dimensionality

Chess

Image processing

$M^7$

$M^{64}$

$M^{nxn}!!!$
Solution: find structures in data
Shared challenge in ML and Cognition

Too much data...
Shared challenge in ML and Cognition

Too few data...
What is Machine Learning?

• “The goal of machine learning is to build computer systems that can adapt and learn from their experience.”
  – Tom Dietterich
A Generic System

Input Variables: $\mathbf{x} = \left( x_1, x_2, \ldots, x_N \right)$

Hidden Variables: $\mathbf{h} = \left( h_1, h_2, \ldots, h_K \right)$

Output Variables: $\mathbf{y} = \left( y_1, y_2, \ldots, y_L \right)$

Parameters: $\mathbf{w} = \left( w_1, w_2, \ldots, w_M \right)$

Training examples: $\left( \mathbf{x}^1, \mathbf{t}^1 \right), \left( \mathbf{x}^2, \mathbf{t}^2 \right), \ldots, \left( \mathbf{x}^D, \mathbf{t}^D \right)$
Handwritten Digit Recognition

\[
y = f_w(x, h)
\]

\(x\): pixelized or pre-processed image.
\(y\): digit class.
\(h\): left/right handed.

\(W\): parameters.
Different types of learning

• **Supervised learning:**
  1. Classification (discrete y),
  2. Regression (continuous y).

• **Unsupervised learning** (no t).
  1. Clustering (h = different groups of types of data).
  2. Density estimation (h = parameters of probability dist.)
  3. Reduction (h= a few latent variable describing high dimensional data).

• **Reinforcement learning** (y = actions, rewards).
Polynomial Curve Fitting

\[ y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j \]
Sum-of-Squares Error Function

\[ E(w) = \frac{1}{2} \sum_{n=1}^{N} \left( y(x_n, w) - t_n \right)^2 \]
0\textsuperscript{th} Order Polynomial

\[ M = 0 \]
$1^{st}$ Order Polynomial

$M = 1$
$3^{rd}$ Order Polynomial

\[ M = 3 \]
$9^{\text{th}} \text{ Order Polynomial}$

\[ M = 9 \]
Over-fitting

Root-Mean-Square (RMS) Error: $E_{\text{RMS}} = \sqrt{2E(w^*)/N}$
## Polynomial Coefficients

<table>
<thead>
<tr>
<th></th>
<th>$M = 0$</th>
<th>$M = 1$</th>
<th>$M = 3$</th>
<th>$M = 9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_0^*$</td>
<td>0.19</td>
<td>0.82</td>
<td>0.31</td>
<td>0.35</td>
</tr>
<tr>
<td>$w_1^*$</td>
<td></td>
<td>-1.27</td>
<td>7.99</td>
<td>232.37</td>
</tr>
<tr>
<td>$w_2^*$</td>
<td></td>
<td></td>
<td>-25.43</td>
<td>-5321.83</td>
</tr>
<tr>
<td>$w_3^*$</td>
<td></td>
<td></td>
<td>17.37</td>
<td>48568.31</td>
</tr>
<tr>
<td>$w_4^*$</td>
<td></td>
<td></td>
<td></td>
<td>-231639.30</td>
</tr>
<tr>
<td>$w_5^*$</td>
<td></td>
<td></td>
<td></td>
<td>640042.26</td>
</tr>
<tr>
<td>$w_6^*$</td>
<td></td>
<td></td>
<td></td>
<td>-1061800.52</td>
</tr>
<tr>
<td>$w_7^*$</td>
<td></td>
<td></td>
<td></td>
<td>1042400.18</td>
</tr>
<tr>
<td>$w_8^*$</td>
<td></td>
<td></td>
<td></td>
<td>-557682.99</td>
</tr>
<tr>
<td>$w_9^*$</td>
<td></td>
<td></td>
<td></td>
<td>125201.43</td>
</tr>
</tbody>
</table>
Data Set Size:

9th Order Polynomial

![Graph of 9th Order Polynomial](image)
Data Set Size:

9th Order Polynomial

$N = 100$
Regularization

- Penalize large coefficient values

\[ \tilde{E}(w) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, w) - t_n\}^2 + \frac{\lambda}{2} \|w\|^2 \]
Regularization: \[ \ln \lambda = -18 \]
Regularization:

\[ \ln \lambda = 0 \]
Regularization: $\ln \lambda$ vs. $E_{\text{RMS}}$
# Polynomial Coefficients

<table>
<thead>
<tr>
<th></th>
<th>$\ln \lambda = -\infty$</th>
<th>$\ln \lambda = -18$</th>
<th>$\ln \lambda = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_0^*$</td>
<td>0.35</td>
<td>0.35</td>
<td>0.13</td>
</tr>
<tr>
<td>$w_1^*$</td>
<td>232.37</td>
<td>4.74</td>
<td>-0.05</td>
</tr>
<tr>
<td>$w_2^*$</td>
<td>-5321.83</td>
<td>-0.77</td>
<td>-0.06</td>
</tr>
<tr>
<td>$w_3^*$</td>
<td>48568.31</td>
<td>-31.97</td>
<td>-0.05</td>
</tr>
<tr>
<td>$w_4^*$</td>
<td>-231639.30</td>
<td>-3.89</td>
<td>-0.03</td>
</tr>
<tr>
<td>$w_5^*$</td>
<td>640042.26</td>
<td>55.28</td>
<td>-0.02</td>
</tr>
<tr>
<td>$w_6^*$</td>
<td>-1061800.52</td>
<td>41.32</td>
<td>-0.01</td>
</tr>
<tr>
<td>$w_7^*$</td>
<td>1042400.18</td>
<td>-45.95</td>
<td>-0.00</td>
</tr>
<tr>
<td>$w_8^*$</td>
<td>-557682.99</td>
<td>-91.53</td>
<td>0.00</td>
</tr>
<tr>
<td>$w_9^*$</td>
<td>125201.43</td>
<td>72.68</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Probability Theory

Apples and Oranges
Probability Theory

Joint Probability

\[ p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} \]

Marginal Probability

\[ p(X = x_i) = \frac{c_i}{N}. \]

Conditional Probability

\[ p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}. \]
Probability Theory

Sum Rule

\[ p(X = x_i) = \frac{c_i}{N} = \frac{1}{N} \sum_{j=1}^{L} n_{ij} \]

\[ = \sum_{j=1}^{L} p(X = x_i, Y = y_j) \]

Product Rule

\[ p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \]

\[ = p(Y = y_j | X = x_i) p(X = x_i) \]
The Rules of Probability

- Sum Rule
  \[ p(X) = \sum_Y p(X, Y) \]

- Product Rule
  \[ p(X, Y) = p(Y|X)p(X) \]
Bayes’ Theorem

\[ p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \]

\[ p(X) = \sum_Y p(X|Y)p(Y) \]

posterior \( \propto \) likelihood \( \times \) prior
Probability Densities

\[ p(x \in (a, b)) = \int_{a}^{b} p(x) \, dx \]

\[ P(z) = \int_{-\infty}^{\infty} p(x) \, dx \]

\[ p(x) \geq 0 \quad \int_{-\infty}^{\infty} p(x) \, dx = 1 \]
Expectations

\[ \mathbb{E}[f] = \sum_{x} p(x)f(x) \]

\[ \mathbb{E}[f] = \int p(x)f(x) \, dx \]

\[ \mathbb{E}_x[f|y] = \sum_{x} p(x|y)f(x) \]

Conditional Expectation (discrete)

\[ \mathbb{E}[f] \approx \frac{1}{N} \sum_{n=1}^{N} f(x_n) \]

Approximate Expectation (discrete and continuous)
Variances and Covariances

\[ \text{var}[f] = \mathbb{E} \left[ (f(x) - \mathbb{E}[f(x)])^2 \right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 \]

\[
\text{cov}[x, y] = \mathbb{E}_{x,y} \left\{ x - \mathbb{E}[x]\right\} \left\{ y - \mathbb{E}[y]\right\} \\
= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]
\]

\[
\text{cov}[x, y] = \mathbb{E}_{x,y} \left\{ x - \mathbb{E}[x]\right\} \left\{ y^T - \mathbb{E}[y^T]\right\} \\
= \mathbb{E}_{x,y}[xy^T] - \mathbb{E}[x]\mathbb{E}[y^T]
\]
The Gaussian Distribution

\[ \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\} \]

\[ \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) \, dx = 1 \]
Gaussian Mean and Variance

\[ \mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) \, x \, dx = \mu \]

\[ \mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) \, x^2 \, dx = \mu^2 + \sigma^2 \]

\[ \text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2 \]
The Multivariate Gaussian

\[ \mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\} \]
Determinant

Determinant of a matrix: $|A|$

See:

http://www.mathsisfun.com/algebra/matrix-determinant.html

Or: ... type det(A) in matlab, like I always do 😊
Curve Fitting Re-visited

\[ y(x_0, w) \]

\[ p(t|x_0, w, \beta) = \mathcal{N}(t|y(x_0, w), \beta^{-1}) \]
Math reminder

Logs transform products into sums:

\[
\ln(a_1a_2...a_K) = \ln\left(\prod_{k=1}^{K} a_k\right) = \sum_{k=1}^{K} \ln(a_k)
\]

Exp is the inverse of log, transforms sums into products:

\[
\ln\left(\exp(x)\right) = x \\
\exp\left(\sum_{k=1}^{K} a_k\right) = \prod_{k=1}^{K} \exp(a_k)
\]
Maximum Likelihood

\[ p(t|x, w, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n | y(x_n, w), \beta^{-1}) \]

\[ \ln p(t|x, w, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, w) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) \]

\[ = \beta E(w) \]

Determine \( w_{\text{ML}} \) by minimizing sum-of-squares error, \( E(w) \).

\[ \frac{1}{\beta_{\text{ML}}} = \frac{1}{N} \sum_{n=1}^{N} \{y(x_n, w_{\text{ML}}) - t_n\}^2 \]
Predictive Distribution

\[ p(t|x, w_{ML}, \beta_{ML}) = \mathcal{N}(t|y(x, w_{ML}), \beta_{ML}^{-1}) \]
MAP: A Step towards Bayes

\[ p(w|\alpha) = \mathcal{N}(w|0, \alpha^{-1}I) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp \left\{ -\frac{\alpha}{2}w^T w \right\} \]

\[ p(w|x, t, \alpha, \beta) \propto p(t|x, w, \beta)p(w|\alpha) \]

\[ \beta \tilde{E}(w) = \frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, w) - t_n \}^2 + \frac{\alpha}{2}w^T w \]

Determine \( w_{MAP} \) by minimizing regularized sum-of-squares error, \( \tilde{E}(w) \).
Bayesian Predictive Distribution

\[ p(t | x, x, t) = \int \int \ldots \int p(t | x, x, t, w) p(w | x, t) dw_1 \ldots dw_N \]

\[ p(t|x, x, t) = \mathcal{N}(t|m(x), s^2(x)) \]
Bayesian perception
- All of our decisions are subject to uncertainty
Ambiguities
Perception as estimation

$\hat{x} = f(s)$
Perception as Bayesian Inference

\[ P(x) \quad \xrightarrow{\text{Likelihood}} \quad P(s \mid x) \]

\[ P(x \mid s) = \frac{P(s \mid x)P(x)}{P(s)} \]

Prior

Likelihood

CNS

Posterior

Model
Ex: Aperture problem

Direction of motion

Sensory input
Ex: Aperture problem

$P(s | x)$

Likelihood

$P(x)$

Prior

$P(x | s)$

Posterior

$X$

Direction

S

Sensory input
Integrating from multiple edges

\[ p(s_1, s_2 | x) = P(s_1 | x) P(s_2 | x) \]

\[ P(x | s_1, s_2) = \frac{1}{Z} p(s_1 | x) p(s_2 | x) \]
Integrating from multiple edges

\[ P(s_1 | x) \quad \text{and} \quad P(s_2 | x) \]

\[ P(x) \quad \text{and} \quad P(x | s) \]
Integrating from multiple edges

\[ P(s_1 | x) \]

\[ P(s_2 | x) \]

\[ P(x) \]

\[ P(x | s) \]
Example of illusions explained by the prior on low velocity

Velocity perception as bayesian inference:
Weiss, Simoncelli and Adelson, 2002
Use of a prior in sensorimotor control

Kording and Wolpert, Nature, 2004
Experimental results

Prediction:

Experimental result:
Cue combination

Ex: The ventriloquism effect
Combining cues from several modalities

\[ p(x | s_a, s_v) = \frac{P(x)P(s_a | x)P(s_v | x)}{Z} \]
Combining cues from several modalities: Gaussian distributions

\[
p(s_a | x) = \frac{1}{\sqrt{2\pi} \sigma_a} \exp\left(\frac{- (x - x_a)^2}{2\sigma_a^2}\right)
\]

\[
p(s_v | x) = \frac{1}{\sqrt{2\pi} \sigma_v} \exp\left(\frac{- (x - x_v)^2}{2\sigma_v^2}\right)
\]

\[
p(x | s_a, s_v) = \frac{1}{\sqrt{2\pi} \sigma_b} \exp\left(\frac{- (x - x_b)^2}{2\sigma_b^2}\right)
\]
Combining cues from several modalities, Gaussian distributions

\[ x_b = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_v^2} x_v + \frac{\sigma_v^2}{\sigma_a^2 + \sigma_v^2} x_a \]

Visual weight  Visual cue  Auditory weight  Auditory cue
Analogy with the center-of-mass

Visual more reliable: Visual capture

Bimodal

Visual

x_{vis}

Auditory

x_{aud}

Bimodal

Auditory more reliable: Auditory capture:

Bimodal

Visual

x_{vis}

Auditory

x_{aud}
Ernst and Banks, 2002

From: Banks et al, Nature, 2002
Temporal predictions

Forward prediction

Hidden state

cause

Observations

\[
p(x_{t+1} | s_{0 \rightarrow t+1}) = \frac{p(s_{t+1} | x_{t+1}) \sum_{x_t} p(x_{t+1} | x_t) p(x_t | s_{0 \rightarrow t})}{Z}
\]
Gaussian distribution: Kalman filter

\[ \hat{x}_{t+1} = (1 - k_t) \hat{x}_t^f + k_t x_t^s \]

\[ k_t = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_t^2} \]

Kalman gain
Ex: Optimal motor control with gaussian distributions

\[
\hat{x}(t+1) = \left(1 - K(t)\right) \hat{x}_P(t) + K(t) x_s(t)
\]

Gharahmani and Wolpert, 2002

Reliability of internal estimates versus sensory feedback
Experimental evidence: Sensorimotor control

Purely sensory feedback

Purely forward estimate (no sensory feedback)

Optimal kalman filter

Wolpert et al, 1995
Control of saccade sequences

\[ \hat{x}_{t+1} = \left(1 - k_t\right) \hat{x}_t^f + k_t x_t^s \]

\[ k_t = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_t^2} \]

Munuera, Morel and Deneve, J Neurosci 2009
Experimental evidence: saccadic eye movements

\[ k_t = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_t^2} \]
Bayesian perception: multiple causes

Infering 3D structures from 2D images.
Knill and Richards, 1996

\[ p(x_1 | s) = \frac{p(x_1) \sum_{x_2} p(s | x_1, x_2) p(x_2)}{Z} \]
Bayesian perception and action: multiple causes

Explaining away

\[ p(x_1 | s) = \frac{p(x_1) \sum_{x_2} p(s | x_1, x_2) p(x_2)}{Z} \]
Example: Cues should be integrated only when they have the same source.

Kording et al, PloS 2007

«Strong fusion»

«Weak fusion»

Common source likely

Common source unlikely
Cues should be integrated only when they have the same source.

Kording et al, PloS 2007

Explaining away

“Same source”

Attraction

“Different source”

Repulsion

“Different source”
Experimental results

Kording et al, PloS 2007