ECE285/SIO209, Machine learning for physical applications,
Spring 2017

Peter Gerstoft, 534-7768, gerstoft@ucsd.edu
We meet Wednesday from 5 to 6:20pm in Spies Hall 330

Text Bishop

Grading  A or maybe S

Classes
First 4 weeks: Focus on theory/implementation
Middle 3 weeks: 50% Applications plus  50% theory
Last 3 weeks: 30% Final Project, 30% Applications plus  50% theory

Applications
Graph theory for source localization: Gerstoft
Source tracking in ocean acoustics: Grad Student Emma Reeves
Aramco Research: Weichang Li, Group leader
Seismic network using mobile phones: Berkeley
Eric Orenstein: identifying plankton
Plus more
- Dictionary learning
-

Homework: Both matlab/python will be used, Just email the code to me (I dont need anything else).
Homework is due 11am on Wednesday. That way we can discuss in class.
Hw 1:

matlab/ python/ ipython/ jupyter ?

Tritoned? https://tritoned.ucsd.edu/
Machine Learning for Geophysical Applications
Peter Gerstoft
noiselab.ucsd.edu

Plan
Unsupervised source localization (graph theory)
Supervised source localization (neural network)
Unsupervised dictionary learning for sound speed

Murphy: “This books adopts the view that the best way to make machines that can learn from data is to use the tools of probability theory, which has been the mainstay of statistics and engineering for centuries.”
Neural Networks (TensorFlow) for tracking a ship

Niu and Gerstoft JASA, 2016
**Input:** Sample cov. matrix: 272 Neurons (16*17/2*2) per frequency at each range

**Output:** binary range vector: 0.1-3km, 138 neurons

Just one middle layer 128 Neurons
TensorFlow implementation

\[ z = f(Vx) \]
\[ y = h(Wz) \]

h: softmax function:
\[ a = Wz \]
\[ y_k = \frac{e^{a_k}}{\sum_j e^{a_j}} \]
FIG. 12. (Color online) Range predictions on training data (a, b, c, d, first row), Test-Data-1 (e, f, g, h, second row) and Test-Data-2 (i, j, k, l, third row) by FNN with multi-frequency inputs. (a)(e)(i) 450, 490, 520, 550 Hz. (b)(f)(j) 560, 590, 620, 650 Hz. (c)(g)(k) 660, 690, 720, 750 Hz. (d)(h)(l) 450, 600, 750, 900 Hz. The time index increment is 10 s for training and Test-Data-1, and 5 s for Test-Data-2.
TRANSFER LEARNING AND DEEP FEATURE EXTRACTION FOR PLANKTONIC IMAGE DATA SETS

Eric C. Orenstein¹ and Oscar Beijbom²

¹ Scripps Institution of Oceanography – University of California San Diego
² Department of Electrical Engineering and Computer Science – University of California Berkeley
IEEE Winter Conference on Applications of Computer Vision 2017 – Paper ID #313
Qingkai Kong is from Berkeley, I have 3GB of data and examples of analysis by students there.
Why we got interested in traffic

March 5—12, 2011
Noise Tracking of Cars/Trains/Airplanes

5200 element Long Beach array (Dan Hollis)
Noise Tracking of Cars/Trains/Airplanes

March 7\textsuperscript{th}, 6-7am, rush hour, Blue Line

Accelerating airplane on Long Beach Airport runway, moving northwest and taking off at about 120 mi/h.

Riahi, Gerstoft, GRL 2015
Thought experiment: Party at a detection array

30-microphone array

100 m
Spectral coherence between $i$ and $j$

$$\hat{C}_{ij}(f) = \frac{1}{N} \sum_{t=1}^{N} X_i(f, t) \cdot \bar{X}_j(f, t)$$

(Normalization: $|X(f, t)|^2 = 1$)
**Objective**

Find coherent but very localized event in a large array. Don’t assume anything about the medium.

**Approach**

Construct network using pair-wise sensor coherence. Exploit network structure to identify sources.
What is Machine Learning?

Many related terms:

• Pattern Recognition
• Neural Networks
• Data Mining
• Adaptive Control
• Statistical Modelling
• Data analytics / data science
• Artificial Intelligence

Machine Learning
Learning: The view from different fields

- **Engineering:** signal processing, system identification, adaptive and optimal control, information theory, robotics, ...

- **Computer Science:** Artificial Intelligence, computer vision, information retrieval, ...

- **Statistics:** learning theory, data mining, learning and inference from data, ...

- **Cognitive Science and Psychology:** perception, movement control, reinforcement learning, mathematical psychology, computational linguistics, ...

- **Computational Neuroscience:** neuronal networks, neural information processing, ...

- **Economics:** decision theory, game theory, operational research, ...

**Physical science is missing!**
ML cannot replace physical understanding.
It might improve or find additional trends

**Machine learning** is interdisciplinary focusing on both mathematical foundations and practical applications of systems that learn, reason and act.
Probabilistic Modelling

• A model describes data that one could observe from a system

• If we use the mathematics of probability theory to express all forms of uncertainty and noise associated with our model...

• ...then inverse probability (i.e. Bayes rule) allows us to infer unknown quantities, adapt our models, make predictions and learn from data.
Bayes Rule

\[ P(\text{hypothesis}|\text{data}) = \frac{P(\text{data}|\text{hypothesis})P(\text{hypothesis})}{P(\text{data})} \]

- Bayes rule tells us how to do inference about hypotheses from data.
- Learning and prediction can be seen as forms of inference.
Some Canonical Machine Learning Problems

- Linear Classification
- Polynomial Regression
- Clustering with Gaussian Mixtures (Density Estimation)
**Linear Classification**

**Data:** \( D = \{(x^{(n)}, y^{(n)})\} \) for \( n = 1, \ldots, N \) data points

- \( x^{(n)} \in \mathbb{R}^D \)
- \( y^{(n)} \in \{+1, -1\} \)

**Model:**

\[
P(y^{(n)} = +1|\theta, x^{(n)}) = \begin{cases} 
1 & \text{if } \sum_{d=1}^{D} \theta_d x_d^{(n)} + \theta_0 \geq 0 \\
0 & \text{otherwise}
\end{cases}
\]

**Parameters:** \( \theta \in \mathbb{R}^{D+1} \)

**Goal:** To infer \( \theta \) from the data and to predict future labels \( P(y|\mathcal{D}, x) \)
Polynomial Regression

Data: \( \mathcal{D} = \{(x^{(n)}, y^{(n)})\} \) for \( n = 1, \ldots, N \)

\[
\begin{align*}
x^{(n)} & \in \mathbb{R} \\
y^{(n)} & \in \mathbb{R}
\end{align*}
\]

Model:

\[
y^{(n)} = a_0 + a_1 x^{(n)} + a_2 x^{(n)^2} + \ldots + a_m x^{(n)^m} + \epsilon
\]

where

\[
\epsilon \sim \mathcal{N}(0, \sigma^2)
\]

Parameters: \( \theta = (a_0, \ldots, a_m, \sigma) \)

Goal: To infer \( \theta \) from the data and to predict future outputs \( P(y|\mathcal{D}, x, m) \)
Clustering with Gaussian Mixtures  
(Density Estimation)

**Data:** \( \mathcal{D} = \{ \mathbf{x}^{(n)} \} \) for \( n = 1, \ldots, N \)

\[ \mathbf{x}^{(n)} \in \mathbb{R}^D \]

**Model:**

\[ \mathbf{x}^{(n)} \sim \sum_{i=1}^{m} \pi_i \, p_i(\mathbf{x}^{(n)}) \]

where

\[ p_i(\mathbf{x}^{(n)}) = \mathcal{N}(\mu^{(i)}, \Sigma^{(i)}) \]

**Parameters:** \( \theta = ((\mu^{(1)}, \Sigma^{(1)}) \ldots, (\mu^{(m)}, \Sigma^{(m)}), \pi) \)

**Goal:** To infer \( \theta \) from the data, predict the density \( p(\mathbf{x}|\mathcal{D}, m) \), and infer which points belong to the same cluster.
Bayesian Modelling

Everything follows from two simple rules:

**Sum rule:** \[ P(x) = \sum_y P(x, y) \]

**Product rule:** \[ P(x, y) = P(x)P(y|x) \]

Prediction:
\[ P(x|D, m) = \int P(x|\theta, D, m)P(\theta|m)d\theta \]

Model Comparison:
\[ P(m|D) = \frac{P(D|m)P(m)}{P(D)} \]
\[ P(D|m) = \int P(D|\theta, m)P(\theta|m) d\theta \]
ML overview

• Output: $y(x)$
  images $x$

**Target vector $y$ or $t$**

• Learning/ training [$x_1$...]
• Test set
• Feature extraction
• Supervised learning--- Making predictions
  – Classification
  – Regression
• Unsupervised learning
  – Clustering
  – Density estimation
• Reinforcement learning
  – Exploration $<$ exploitation

Nmist data set

0 1 2 3 4
5 6 7 8 9
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

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

$M = 0$

Graph showing a 0th order polynomial with points plotted and a horizontal line at $M = 0$. The graph includes axes labeled $x$ and $t$. The plot shows the polynomial curve and the line intersecting at $M = 0$.
1\textsuperscript{st} Order Polynomial

\[ M = 1 \]
$3^{rd}$ Order Polynomial

![Graph showing a 3rd order polynomial with points and curves](image)
9\textsuperscript{th} Order Polynomial

\[ M = 9 \]
Over-fitting

Root-Mean-Square (RMS) Error: $E_{RMS} = \sqrt{2E(w^*)/N}$
### 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>-1.27</td>
<td>7.99</td>
<td></td>
<td>232.37</td>
</tr>
<tr>
<td>$w_2^*$</td>
<td></td>
<td>-25.43</td>
<td>-5321.83</td>
<td></td>
</tr>
<tr>
<td>$w_3^*$</td>
<td></td>
<td>17.37</td>
<td>48568.31</td>
<td></td>
</tr>
<tr>
<td>$w_4^*$</td>
<td></td>
<td></td>
<td>-231639.30</td>
<td></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

\[ N = 15 \quad \text{and} \quad N = 100 \]
Regularization

- Penalize large coefficient values

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

\[ \ln \lambda = -18 \]

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

<table>
<thead>
<tr>
<th>$w_i^*$</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

- **Marginal Probability**
  \[ p(X = x_i) = \frac{c_i}{N} . \]

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

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

Product Rule

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

\[ = p(Y = y_j \mid X = x_i)p(X = x_i) \]

Sum Rule

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

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

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

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

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

\[ p_y(y) = p_x(x) \left| \frac{dx}{dy} \right| = p_x(g(y)) \left| g'(y) \right| \]
Expectations

\[ \mathbb{E}[f] = \sum_x p(x) f(x) \hspace{2cm} \mathbb{E}[f] = \int p(x) f(x) \, dx \]

\[ \mathbb{E}_x[f|y] = \sum_x p(x|y) f(x) \hspace{2cm} \text{Conditional Expectation (discrete)} \]

\[ \mathbb{E}[f] \approx \frac{1}{N} \sum_{n=1}^N f(x_n) \hspace{2cm} \text{Approximate Expectation (discrete and continuous)} \]

Variances and Covariances

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

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

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

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

\[ \mathcal{N}(x|\mu, \sigma^2) > 0 \]

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

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

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

\[ p(x|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2) \]
Maximum (Log) Likelihood

\[ \ln p(x | \mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi) \]

\[ \mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n \]
\[ \sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ML})^2 \]
Curve Fitting Re-visited

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

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

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

\[ \beta E(w) \]

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

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

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

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

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

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

Determine \(w_{\text{MAP}}\) by minimizing regularized sum-of-squares error, \(\tilde{E}(w)\).
Minimum Misclassification Rate

\[ p(\text{mistake}) = p(x \in \mathcal{R}_1, \mathcal{C}_2) + p(x \in \mathcal{R}_2, \mathcal{C}_1) \]

\[ = \int_{\mathcal{R}_1} p(x, \mathcal{C}_2) \,dx + \int_{\mathcal{R}_2} p(x, \mathcal{C}_1) \,dx. \]
Reject Option

The diagram illustrates the probability density functions $p(C_1|x)$ and $p(C_2|x)$ as functions of $x$. The reject region is defined by the threshold $\theta$, and observations falling within this region are rejected.
OLD