Note:

- HW#5 submission closed now.
- All homework are completed now 😊

Note:

- Continue working on your lit. survey project.
- Project report due in three weeks
  - Read the assignment thoroughly
  - Late policy will apply.
- Project presentation in three weeks
  - Submit your slides two days prior (5/12, 5pm) by email
  - Read the assignments again to refresh your mem.
  - More details for presentation will be emailed soon
  - Presentation in alphabetical order
Regression & Learning

CSC 872
Pattern Analysis and Machine Intelligence

What is Regression?

What is the population in year 2025?

1. Fit a line
2. Find our prediction

- Regression is a statistical analysis to find a function representing input-output relation from data samples

http://www.populationmedia.org/issues/pogrowth_data.html
What is Regression?

- We must choose appropriate function form to do regression.

What is the population in year 2025?

1. Fit a curve
2. Find our prediction

In Nutshell

- Goal: estimate input-output relation from data
- For: prediction/forecasting/modeling
- You need to
  
1) Pick a form of parametric function
   - Line:
   - Polynomial Curves:
   - General Curves:
   - General Functions:

2) Fit the function to the data
   - Maximum likelihood estimation (MLE) is the foundation
Learning Machine Interpretation

• Learning Machine: $y = f(x)$

• Input: independent variable
  – E.g., $x =$ year

• Output: dependent variable
  – E.g., $y =$ population

• Function: parameterized by $W$
  – E.g., $f(x,w) = wx$: line

\[ y_{\text{est}} = Wf(x) \]

Training Data $X$ (Supervised ML)

• Supervised learning

\[ y_i = \begin{cases} 
+1 & \text{if bear} \\
-1 & \text{otherwise} 
\end{cases} \]

\[ X = \{(x_i, y_i)\}^N_{i=1} \]

\[ y = f(x_1) \]

\[ x_2 = g(x_1) \]
### Regression Types

Depends on the form of parameterized functions

- Linear Regression (line/plane/hyperplane)
- Polynomial Regression (polynomial curve)
- Non-linear Regression (general curve)
- Radial-Basis Function Regression (basis sum)
- Piecewise Linear Regression (line segments)
- Non-parametric Regression (KDE)
- Robust Regression (robust estimation!!!)
**Linear Regression**

- **Simplest one parameter case**
  \[ y \in \mathbb{R}, x \in \mathbb{R}, w \in \mathbb{R}. \]

- Data is formed by: \( y = wx + \text{noise} \)
  - Unknown scalar \( w \)
  - Noise is independent random variable
  - Noise is normally-distributed with zero mean & \( \sigma^2 \)

- Output \( y \) is then also a random variable with
  \[
P(y|x, w) = \text{Normal}( \text{mean } \langle wx \rangle, \text{variance } \sigma^2)\]

- Given data: \( N \) i.i.d. evidences \((x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\)
- Regression Problem: Find \( w \) from data such that ....

---

**Bayesian Linear Regression**

- **Find \( w \) from data such that it maximizes the posterior distribution:**
  \[
P( w | (x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N) )
  \]

- Infer \( P(w|data) \) from data likelihood \( P(y|x, w) \) using Bayes rules!
  - Conjugate prior etc, A bit complicated so...
Maximum-Likelihood Estimate

- Find $w$ from data such that it maximizes the data likelihood function:

$$P(y \mid x_1, x_2, \ldots, x_N, w) = N(y; wx, \sigma^2)$$

- As usual, let’s do some algebra to simplify

**Algebra Joy: You know this by now**

- For what $w$ is this data most likely to have happened?
- For what $w$, is $P(y_1, \ldots, y_N \mid x_1, \ldots, x_N, w)$ maximized?
- For what $w$, is $\prod_{i=1}^N P(y_i \mid x_i, w)$ maximized?
- For what $w$, is $\prod_{i=1}^N \exp\left(-\frac{1}{2}(y_i-wx_i)^2\right)$ maximized?
- For what $w$, is $\sum_{i=1}^N \frac{1}{2}(y_i-wx_i)^2$ maximized? This is known as Least Squares method.
- For what $w$, is $\sum_{i=1}^N (y_i - wx_i)^2$ minimized?
**Least Squares Method**

- MLE of $w$ is one that minimizes the sum-of-squares of residuals (errors)

$$E(w) = \sum_i (y_i - wx_i)^2$$

**Really a quadratic optimization**

- MLE of $w$ is one that minimizes the sum-of-squares of residuals (errors)

$$E(w) = \sum_i (y_i - wx_i)^2 = \sum_i (x_i^2w^2 - 2wx_iyw_i + y_i^2)$$

$$= \sum_i x_i^2w^2 - 2\sum_i x_iyw_i + \sum_i y_i^2$$

- We want to minimize a quadratic function of $w$

$$\frac{\partial E(w)}{\partial w} = 0$$
STOP: We have a closed-form solution!

- For linear regression with normal-distributed noise
- **MLE = Least Squares !!!**

\[ \hat{w} = \arg\min_w E(w) = \arg\min_w \sum_i (y_i - w x_i)^2 \]

\[ \Leftrightarrow \frac{\partial E(w)}{\partial w} = 0 \]

\[ w = \frac{\sum_i x_i y_i}{\sum_i x_i^2} \]

- \( r \): Pearson correlation coefficient
- \( \sigma_x \): standard deviation of \( \{x_i\} \)
- \( \sigma_y \): standard deviation of \( \{y_i\} \)

\[ y = w x = \frac{\sum_i x_i y_i}{\sum_i x_i^2} x = \frac{\sigma_y}{\sigma_x} r x \]

**Multivariate Case?**

- What if input \( x \) is a vector \((x_1, \ldots, x_M)^T\)?
- Model is \( y = w^T x + \epsilon \)
  \[ = w_1 x_1 + w_2 x_2 + \cdots + w_M x_M + \cdots + w_M x_M + \epsilon \]
- Given data: \((x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\)

\[ X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \]

- MLE of \( w \) is given by
  \[ w = (X^T X)^{-1} X^T Y \] (Pseudo Inverse)

\[ y = w^T x = ((X^T X)^{-1} X^T Y)^T x \]
Constant Term?

• What if the line does not intersect the origin?

\[
Y = w_1 x_1 + w_2 x_2 + \epsilon
\]

Model is

\[
y = w^t z + \epsilon, \quad z = (1, x^t) \]

\[
y = w_0 + w_1 x_1 + w_2 x_2 + \epsilon
\]

Given data: \((x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\)

\[
Z = \begin{bmatrix}
    x_1^t \\
    x_2^t \\
    \vdots \\
    x_N^t
\end{bmatrix} = \begin{bmatrix}
    (1, x_1^t) \\
    (1, x_2^t) \\
    \vdots \\
    (1, x_N^t)
\end{bmatrix}
\]

MLE of \(w\) is given by

\[
w = (Z^t Z)^{-1} Z^t Y
\]

Most general linear regression formula!!

\[
y = w^t z = ((Z^t Z)^{-1} Z^t Y) \begin{bmatrix} 1 \\ x \end{bmatrix}
\]
Polynomial Regression

• What if I want to fit a polynomial curve?

• You can reuse the linear formula!!!

Quadratic Regression

• Model for 2D is

\[ y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_1 x_2 + w_5 x_2^2 + \epsilon \]

\[ = w^T z + \epsilon, \quad z = (1, x_1, x_2, x_1^2, x_1 x_2, x_2^2) \]

• Given data: \((x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N), x_i = (x_{i1},x_{i2})\)

\[ Z = \begin{bmatrix} z_1^T \\ z_2^T \\ \vdots \\ z_N^T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & x_{11}^2 & x_{11} x_{12} & x_{12}^2 \\ x_{21} & x_{22} & x_{21}^2 & x_{21} x_{22} & x_{22}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & x_{N1}^2 & x_{N1} x_{N2} & x_{N2}^2 \end{bmatrix} \]

\[ Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \]

• MLE of \(w\) is given by

\[ w = (Z^T Z)^{-1} Z^T Y \]

But all these are same

\[ y = w^T z = ((Z^T Z)^{-1} Z^T Y) z \]
Q\textsuperscript{th} degree Polynomial Regression

- Model is the same but with different \textbf{z}
  
  \[
  y = w^t z + \epsilon, \quad z = (1, q(x))
  \]

- Given data: \((x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\)

- \(q(x)\): all products of powers of inputs up to Q\textsuperscript{th} degree

\[
Z = \begin{bmatrix}
    z_1^t \\
    z_2^t \\
    \vdots \\
    z_N^t \\
\end{bmatrix} = \begin{bmatrix}
    1 & q(x_1) \\
    1 & q(x_2) \\
    \vdots & \vdots \\
    1 & q(x_N) \\
\end{bmatrix}, \quad Y = \begin{bmatrix}
    y_1 \\
    y_2 \\
    \vdots \\
    y_N \\
\end{bmatrix}
\]

- MLE of \(w\) is given by
  \[
  w = (Z^t Z)^{-1} Z^t Y
  \]

\[
y = w^t z = ((Z^t Z)^{-1} Z^t Y)^t z
\]

Radial Basis Function Regression

- Can we generalize the idea of the Polynomial Regression?
  - Basically, you construct \(Z\) with different \(z\) with various products of inputs
  - Then, use the same pseudo inverse formula

- Let’s construct \(z\) with some function \(\phi(x)\) of input \(x\)

- Model is \(y = w_0 + w_1 \phi_1(x) + \cdots + w_K \phi_K(x) + \epsilon\)

- \(B(x)\) is called a \textit{basis} whose linear combination gives an output

- We choose the basis to be symmetric about a center \(c\) with spread \(W\) then call it \textit{radial basis function}

\[
\phi_k(x) = \text{RadialBasisFunction}(\frac{|x-c_k|}{W_k})
\]

- RBF Regression performs the linear regression with \(B(x)\) defined with the radial basis function
Non-linear Regression

- What if I want to fit more general nonlinear function $f(x;w)$?
- Let’s do the same as before!
  - Assume a general model of $y = f(x; w) + \epsilon$
  - Normally-distributed independent noise
  - Likelihood $P(Y|X,w)$ is Normal(mean $f(x;w)$, variance $\sigma^2$)
  - MLE of $w$ = LS of $w$

$$w = \arg\min_w \sum_{i=1}^{N} (y_i - f(x_i; w))^2$$

$$\iff \frac{\partial}{\partial w} \sum_{i} (y_i - f(x_i; w))^2 = 0$$

$$\iff \sum_{i} (y_i - f(x_i; w)) \frac{\partial f(x_i; w)}{\partial w} = 0$$

Ooops! how to solve this about $w$???
We are doomed. we are stuck here?…

Energy Optimization

- Recall the non-parametric modeling lecture…
- The savior is to go “Iterative” to solve $w = \arg\min_w E(w)$
- Minimizing the Energy/Error/Cost/Potential function by
  - Define an iterative step $move(w,E(w))$
  - Then find an initial solution $w_0$
  - Then find a sequence $w_0, w_1, \ldots, w_m$ by doing

  step1: $w_{new} = w_{old} + move(w_{old}, E(w))$

  step2: $w_{old} = w_{new}$

  step3: go to step1

- To do this right, you need to find $move(w,E(w))$
  so that $w_m$ converges to a local minimum of $E(w)$
In general

- You want \( w = \text{argmin}_w E(w) \)

\[ \text{Iterative solution can get stuck at local minimum depending on initialization} \]

- What if I have a maximization problem?

\[ w = \text{argmax}_w E(w) \]

\[ w = \text{argmin}_w -E(w) \]

How are we going to solve it?

- Various ways
  - Line search
  - Hill-Climbing
  - Gradient-Descent (Steepest-Descent/Ascent)
  - Conjugate-Gradient
  - Levenberg-Marquart
  - Newton’s Method
  - Simulated Annealing
  - EM-algorithm
  - Mean Shift
  - More and more...

\[ \text{Optimization} \]

We study this during the lectures for search methods in AI
Gradient-Descent

- We define the step move function as a negative of partial derivatives of the energy w.r.t. the unknown parameter $W$

$$w_{new} = w_{old} - \eta \frac{\partial E(w)}{\partial w} \bigg|_{w=w_{old}}$$

$\eta$ is a learning rate set to a small constant (e.g., 0.05)

Steepest-Descent/Ascent

- Umm, sorry but, I want to maximize really, want to go up
- DON’T LIKE GOING DOWN!!!
- Just flip the sign!!!

$$w_{new} = w_{old} + \eta \frac{\partial E(w)}{\partial w} \bigg|_{w=w_{old}}$$

Gradient-Descent

- Steepest-Ascent: a type of greedy iterative search algorithm we learned in our lecture on search for AI

- Umm: the same
- Gradient-Ascent = Steepest-Ascent
Simulated Annealing

• Are there a way to avoid getting stuck in a local minimum?
• Yes: called "simulated annealing" making it stochastic
  step 1: make a random move
  step 2: take this move if reduces the energy
  step 3: else take it with certain acceptance probability
  step 4: go to step 1

Acceptance Prob. \[ P = \exp\left(\frac{E(w_{\text{old}}) - E(w_{\text{new}})}{T}\right) \]

• When temperature \( T_i \) is
  – High: random walk
  – Low: stochastic steepest descent
• Can converge to the global minimum when scheduling a gradual decreasing of the temperature (cooling schedule)

Summary

• Regression & Learning
  – What is regression?
  – Maximum Likelihood Estimate & Least Squares Method
  – Linear regression
  – Polynomial regression
  – Radial Basis Function regression
  – Gradient-descent
  – Simulated Annealing

• Next
  – Artificial Neural Network
  – End of the LDA FP