CS 4491/CS 7990
SPECIAL TOPICS IN BIOINFORMATICS

* Some contents are adapted from Dr. Hung Huang and Dr. Chengkai Li at UT Arlington

Mington Kang, PhD
Computer Science, Kennesaw State University
Terminology

- **Features**
  - An individual measurable property of a phenomenon being observed
  - The number of features or distinct traits that can be used to describe each item in a quantitative manner
  - May have implicit/explicit patterns to describe a phenomenon
  - E.g., Pixels in images, DNA bases and gene expressions in bioinformatics

- **Samples**
  - Items to process (classify or cluster)
  - Can be a document, a picture, a sound, a video, or a patient

Reference: http://www.slideshare.net/rahuldausa/introduction-to-machine-learning-38791937
Terminology

- **Feature vector**
  - An N-dimensional vector of numerical features that represent some objects
  - A sample consists of feature vectors

- **Feature extraction (feature selection)**
  - Preparation of feature vector
  - Transforms the data in the high-dimensional space to a space of fewer dimensions

Reference: http://www.slideshare.net/rahuldausa/introduction-to-machine-learning-38791937
Examples

Features:
1. Color: Radish/Red
2. Type: Fruit
3. Shape etc...

Features:
1. Sky Blue
2. Logo
3. Shape etc...

Features:
1. Yellow
2. Fruit
3. Shape etc...

Reference: http://www.slideshare.net/rahuldausa/introduction-to-machine-learning-38791937
Features in Bioinformatics

- DNA sequences in sequence alignments
- Number of Minor Alleles in GWAS and eQTL mapping study
- Gene expression in gene regulatory networks
- Protein sequences in protein structures
- Pixels in medical images
Data In Machine Learning

- $x_i$: input vector, independent variable
  
  \[
  x_i = \begin{bmatrix}
  x_{i,1} \\
  x_{i,2} \\
  \vdots \\
  x_{i,n}
  \end{bmatrix}, \quad x_{i,j} \in \mathbb{R}
  \]

- $y$: response variable, dependent variable
  - $y \in \{-1, 1\}$ or $\{0, 1\}$: binary classification
  - $y \in \mathbb{R}$: regression
  - Predict a label when having observed some new $x$
Types of Variable

- **Categorical variable**: discrete or qualitative variables
  - **Nominal**:
    - Have two or more categories, but which do not have an intrinsic order
  - **Dichotomous**
    - Nominal variable which have only two categories or levels.
  - **Ordinal**
    - Have two or more categories, which can be ordered or ranked.
- **Continuous variable**
Mathematical Notation

- Matrix: uppercase bold Roman letter, $\mathbf{X}$
- Vector: lower case bold Roman letter, $\mathbf{x}$
- Scalar: lowercase letter
- Transpose of a matrix or vector: superscript T or ‘$
- E.g.

  - Row vector: $(x_1, x_2, \ldots, x_N)$
  - Corresponding column vector: $\mathbf{x} = (x_1, x_2, \ldots, x_N)^T$
  - Matrix: $\mathbf{X} = \{x_1, x_2, \ldots, x_p\}$
Transpose of a Matrix

- Operator which flips a matrix over its diagonal
  - Switch the row and column indices of the matrix
  - Denoted as $A^T, A’, A^{\text{tr}},$ or, $A^t$.
  $$[A^T]_{ij} = [A]_{ji}$$

- If $A$ is an $m*n$ matrix, $A’$ is an $n*m$ matrix
- $(A^T)^T = A$
- $(A + B)^T = A^T + B^T$
- $(AB)^T = B^T A^T$

\[
A = \begin{bmatrix}
1 & 2 \\
3 & 4 \\
5 & 6
\end{bmatrix}
\]
The inverse of a square matrix $A$, sometimes called a reciprocal matrix, is a matrix $A^{-1}$ such that $AA^{-1} = I$ where $I$ is the identity matrix.

The Inverse of a Matrix is the same idea but we write it $A^{-1}$.
Inverse of a Matrix

Example in 2*2 Matrix

\[
\begin{bmatrix}
a & b \\
c & d
\end{bmatrix}^{-1} = \frac{1}{\text{determinant}} \begin{bmatrix}
d & -b \\
-c & a
\end{bmatrix}
\]

\[
\begin{bmatrix}
4 & 7 \\
2 & 6
\end{bmatrix}^{-1} = \frac{1}{4 \times 6 - 7 \times 2} \begin{bmatrix}
6 & -7 \\
-2 & 4
\end{bmatrix}
\]

\[
= \frac{1}{10} \begin{bmatrix}
6 & -7 \\
-2 & 4
\end{bmatrix}
\]

\[
= \begin{bmatrix}
0.6 & -0.7 \\
-0.2 & 0.4
\end{bmatrix}
\]

Reference: https://www.mathsisfun.com/algebra/matrix-inverse.html
Inverse of a Matrix

- If determinant is zero?

\[
\begin{bmatrix}
3 & 4 \\
6 & 8 \\
\end{bmatrix}^{-1} = \frac{1}{3 \times 8 - 4 \times 6} \begin{bmatrix}
8 & -4 \\
-6 & 3 \\
\end{bmatrix} = \frac{1}{24 - 24} \begin{bmatrix}
8 & -4 \\
-6 & 3 \\
\end{bmatrix}
\]

- We call this matrix “Singular”

Reference: https://www.mathsisfun.com/algebra/matrix-inverse.html
Supervised learning

Data: \( D = \{d_1, d_2, \ldots, d_n\} \) a set of \( n \) samples
where \( d_i = \langle x_i, y_i \rangle \)
\( x_i \) is an input vector and \( y_i \) is a desired output

Objective: learning the mapping \( f: X \to Y \)
subject to \( y_i \approx f(x_i) \) for all \( i = 1, \ldots, n \)

Regression: \( Y \) is continuous
Classification: \( Y \) is discrete
Linear Regression

- Review of Linear Regression in Statistics
Linear Regression

Given examples $(x_i, y_i)_{i=1 \ldots n}$

Predict $y_{n+1}$ given a new point $x_{n+1}$
Linear Regression

\[ \hat{y}_{n+1} \]

\[ x_{n+1} \]

Temperature
Linear Regression

Linear Regression

- How to represent the data as a vector/matrix
  - We assume a model:
    \[ y = b_0 + bX + \epsilon, \]
  - where \( b_0 \) and \( b \) are intercept and slope, known as coefficients or parameters. \( \epsilon \) is the error term (typically assumes that \( \epsilon \sim N(\mu, \sigma^2) \))
Correlation (r)

- Linear association between two variables
- Show how to determine both the nature and strength of relationship between two variables
- Correlation lies between +1 to -1
- Zero correlation indicates that there is no relationship between the variables
- Pearson correlation coefficient
  - most familiar measure of dependence between two quantities
Correlation (r)
Correlation (r)

$$\rho_{X,Y} = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y},$$

where $E$ is the expected value operator, $\text{cov}(,)$ means covariance, and $\text{corr}(,)$ is a widely used alternative notation for the correlation coefficient.

Reference: https://en.wikipedia.org/wiki/Correlation_and_dependence
Coefficient of Determination ($r^2$)

- Coefficient of determination lies between 0 and 1
- Represented by $r^2$
- Measure of how well the regression line represents the data
- If $r = 0.922$, then $r^2 = 0.85$
  - Means that 85% of the total variation in $y$ can be explained by the linear relationship between $x$ and $y$ in linear regression
  - The other 15% of the total variation in $y$ remains unexplained
Linear Regression

- **Simple linear regression**
  - A single independent variable is used to predict

- **Multiple linear regression**
  - Two or more independent variables are used to predict
Linear Regression

- How to represent the data as a vector/matrix
  - Include bias constant (intercept) in the input vector
    - \( X \in \mathbb{R}^{n \times (p+1)} \), \( y \in \mathbb{R}^{n} \), and \( b \in \mathbb{R}^{p+1} \)

\[
X = \{1, x_1, x_2, ..., x_p\}
\]
\[
y = \{y_1, y_2, ..., y_n\}^T
\]
\[
b = \{b_0, b_1, b_2, ..., b_p\}^T
\]
Linear Regression

- Find the optimal coefficient vector $\mathbf{b}$ that makes the most similar observation

$$
\begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{bmatrix} \approx 
\begin{bmatrix}
1 & x_{11} & \cdots & x_{p1} \\
1 & x_{12} & \cdots & x_{p1} \\
\vdots & \vdots & \ddots & \vdots \\
1 & x_{1n} & \cdots & x_{pn}
\end{bmatrix}
\begin{bmatrix}
b_0 \\
b_1 \\
\vdots \\
b_p
\end{bmatrix}
$$

or

$$
\begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{bmatrix} = 
\begin{bmatrix}
1 & x_{11} & \cdots & x_{p1} \\
1 & x_{12} & \cdots & x_{p1} \\
\vdots & \vdots & \ddots & \vdots \\
1 & x_{1n} & \cdots & x_{pn}
\end{bmatrix}
\begin{bmatrix}
b_0 \\
b_1 \\
\vdots \\
b_p
\end{bmatrix} + 
\begin{bmatrix}
e_1 \\
e_2 \\
\vdots \\
e_n
\end{bmatrix}
$$
**Ordinary Least Squares (OLS)**

\[ y = Xb + e \]

- Estimate the unknown parameters \( b \) in linear regression model
- Minimizing the sum of the squares of the differences between the observed responses and the predicted by a linear function

\[
\text{Residual Sum of Squares (RSS)} = \sum_{i=1}^{n} (y_i - x_i b)^2
\]
Ordinary Least Squares (OLS)

\[ X_i = \begin{pmatrix} 1 \\ x_{i,1} \\ x_{i,2} \end{pmatrix} \]
Need to minimize the error

$$\min J(b) = \sum_{i=1}^{n} (y_i - x_i b)^2$$

To obtain the optimal set of parameters ($b$), derivatives of the error w.r.t. each parameter must be zero.
Optimization

\[ J = e^T e = (y - Xb)'(y - Xb) \]
\[ = (y' - b'X')(y - Xb) \]
\[ = y' y - y'Xb - b'X'y + b'X'Xb \]
\[ = y' y - 2b'X'y + b'X'Xb \]

\[ \frac{\partial e'e}{\partial b} = -2X'y + 2X'Xb = 0 \]

\[ (X'X)b = X'y \]
\[ \hat{b} = (X'X)^{-1}X'y \]

Matrix Cookbook: https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf
Assessment of coefficient estimates

Regressions differing in accuracy of prediction.

Assessment of coefficient estimates

- **Standard error of an estimator**
  - Reflects how it varies under repeated sampling
  - \( SE(\hat{b})^2 = \frac{|Y - \hat{Y}|^2}{|x - \hat{x}|^2} \)
  - Where \(|a - b|^2\) represents \( \sum_{i=1}^{n}(a_i - b_i)^2 \)

- **Used to compute confidence intervals**
  - A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.
  - \( \hat{b} \pm 2 \cdot SE(\hat{b}) \)

Reference:
Confidence intervals

- There is approximately a 95% chance that the interval
  \[ \hat{b} - 2 \cdot \text{SE}(\hat{b}), \hat{b} + 2 \cdot \text{SE}(\hat{b}) \]
  will contain the true value of \( b \)

- What happens if the interval includes zero?

Hypothesis testing

- Standard errors can also be used to perform hypothesis tests on the coefficients.
- The most common hypothesis test involves testing
  - the null hypothesis
    - $H_0$ : There is no relationship between $X$ and $Y$
  - versus the alternative hypothesis
    - $H_A$ : There is some relationship between $X$ and $Y$.

## Hypothesis testing

- Mathematically,
  
  \[ H_0: b = 0 \]

  versus

  \[ H_A: b \neq 0 \]

- \( b = 0 \) indicates that the feature does not have relationship with \( y \)

---

Hypothesis testing

- To test the null hypothesis, compute a t-statistic, given by:

\[ t = \frac{\hat{b} - 0}{SE(b)} \]

- This will have a t-distribution with n-2 degrees of freedom, assuming \( b = 0 \).

- Using statistical software, it is easy to compute the probability of observing any value equal to \(|t|\) or larger.

- We call this probability the \( p\)-value

If p-value is less than 0.05 (or 0.01), we consider the feature is statistically significant.

- p < 0.05: significant with 95% confidence
- p < 0.01: significant with 99% confidence

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>7.0325</td>
<td>0.4578</td>
<td>15.36</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>TV</td>
<td>0.0475</td>
<td>0.0027</td>
<td>17.67</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>

Linear regression with categorical variables

- We assumed that all variables are continuous variables

- Categorical variables:
  - Ordinal variables - Encode data with continuous values
    - Evaluation: Excellent (5), Very good (4), Good (3), Poor (2), Very poor (1)
  - Nominal variables – Use dummy variables
    - Department: Computer, Biology, Physics

<table>
<thead>
<tr>
<th></th>
<th>Computer</th>
<th>Biology</th>
<th>Physics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Biology</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Physics</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Linear regression for classification

- For binary classification
  - Encode class labels as $y = \{0, 1\}$ or $\{-1, 1\}$
  - Apply OLS
  - Check which class the prediction is closer to
    - If class 1 is encoded to 1 and class 2 is -1.

  $$\text{class 1} \quad \text{if } f(x) \geq 0$$
  $$\text{class 2} \quad \text{if } f(x) < 0$$

- Logistic regression
  - We will cover this later
Linear regression for classification

- **Multi-label classification**
  - Encode classes label as:

<table>
<thead>
<tr>
<th></th>
<th>Computer</th>
<th>Biology</th>
<th>Physics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Biology</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Physics</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

- Perform linear regression multiple times for each class
- Consider $\mathbf{y}$ and $\mathbf{b}$ as matrix
Feature scaling

- Standardize the range of independent variables (features of data)
- A.k.a Normalization or Standardization
- Better for
  - Regularization
  - Gradient Descent
Standardization or Z-score normalization

- Rescale the data so that the mean is zero and the standard deviation from the mean (standard scores) is one

\[ x_{\text{norm}} = \frac{x - \mu}{\sigma} \]

- \( \mu \) is mean, \( \sigma \) is a standard deviation from the mean (standard score)
Min-Max scaling

- Scale the data to a fixed range – between 0 and 1

\[ x_{\text{norm}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \]
Scaling to unit length

- Scale the data into a unit length vector

\[ x_{\text{norm}} = \frac{x}{\|x\|} \]
Assumptions in Linear regression

- Linearity of independent variable in the predictor
  - normally good approximation, especially for high-dimensional data

- Error has normal distribution, with mean zero and constant variance
  - important for tests

- Independent variables are independent from each other
  - Otherwise, it causes a *multicollinearity* problem; two or more predictor variables are highly correlated.
    - Should remove them
Bias-Variance Tradeoff

- **Bias**: underfitting problem caused by not considering all information of the dataset
- **Variance**: overfitting problem caused by considering all data even including noise and outliers.

https://webdocs.cs.ualberta.ca/~greiner/C-466/SLIDES/3b-Regression.pdf
Different views between Statistics and CS

- In Statistics, description of the model is often more important.
  - Which variables are more informative to describe the responses? → p-values
  - How much information do the variables have?
  - Bioinformatics often emphasize this aspect.

- In Computer Science, the accuracy of prediction and classification is more important.
  - How well can we predict/classify?
Generative vs. Discriminative

- Generative model:
  
  Models how the data is generated in order to categorize a class

- Discriminative model:
  
  Does not care about interpretation of the model, but simply focuses on categorizing a given class