Lecture 1

Math Basics for DS and ML

Scalar-Vector Operations

Consider a scalar $a \in \mathbb{R}$ and a vector $\bar{x} \in \mathbb{R}^{N \times 1}$. Then,

$$a\bar{x} = \begin{bmatrix} ax_1 \\ ax_2 \\ \vdots \\ ax_N \end{bmatrix}$$

For scalar-vector addition, since $\bar{x} + a$ is not valid, we consider

$$\bar{x} + a\mathbb{1} = \begin{bmatrix} x_1 + a \\ x_2 + a \\ \vdots \\ x_N + a \end{bmatrix}$$

Vector-Vector Operations

For vectors $\bar{x}, \bar{y} \in \mathbb{R}^{N \times 1}$,

$$\bar{x} + \bar{y} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_N + y_N \end{bmatrix}$$

Also, we can define an inner product,

$$\bar{x} \cdot \bar{y} = \bar{x}^T \bar{y} = \langle x, y \rangle = x_1 y_1 + \ldots + x_N y_N$$

Similarly,

$$\bar{x} \odot \bar{y} = \begin{bmatrix} x_1 y_1 \\ x_2 y_2 \\ \vdots \\ x_N y_N \end{bmatrix}$$

which is also a $N \times 1$ column vector.

Sub-Spaces

For a vector space V,
If $v_1, v_2 \in V$, then -

1. $cv_i \in V$, where $c \in \mathbb{R}$
2. $v_1 + v_2 \in V$

Now, if $w_1, w_2 \in V$ and $c_1, c_2 \in \mathbb{R}$, then

$$c_1 w_1 + c_2 w_2 \in W \subset V,$$

where W is a subspace described by w_1 and w_2.

Matrix-Matrix Operations

Consider two matrices $X, Y \in \mathbb{R}^{M \times N}$. And column vectors are $x, y \in \mathbb{R}^{N \times 1}$.
Then,

$$X + Y = \begin{bmatrix} x_{11} + y_{11} & \cdots & x_{1N} + y_{1N} \\ \cdots & \cdots & \cdots \\ x_{M1} + y_{M1} & \cdots & x_{MN} + y_{MN} \end{bmatrix}$$

For $Z \in \mathbb{R}^{N \times P}$,

$$X \cdot Z = \begin{bmatrix} x_{11}z_{11} + \ldots + x_{1N}z_{N1} & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & x_{M1}z_{1P} + \ldots + x_{MN}z_{NP} \end{bmatrix}$$

Also,

$$X \odot Y = \begin{bmatrix} x_{11}y_{11} & \cdots & x_{1N}y_{1N} \\ \cdots & \cdots & \cdots \\ x_{M1}y_{M1} & \cdots & x_{MN}y_{MN} \end{bmatrix}$$

Transpose, Determinant and Inverse of a matrix

Let the elements of X^T be b_{ij}. Then,

$$b_{ij} = a_{ji},$$

for all i, j in range and a_{ij} are the elements of X.

$X_{M \times N} X^{-1}_{M \times N} = I_{M \times N}$
If X is invertible $\implies$ X should be Full Rank, where $Rank(X)$ = Number of independent rows and columns of a matrix

Pseudo-Inverse

If, $X^{+}_{N \times M} X_{M \times N} = I_{N \times N}$, then X^{+} is the pseudo-inverse of the non-square matrix X.

The formula for X^+ is,

$$(X^H X)^{-1} X^H,$$

where X^H is the conjugate transpose of X. If all entries of X are real, then X^H becomes X^T (transpose matrix).

Eigen Decomposition

We have a matrix $A(N \times N)$ for which we are trying to find a vector v_i for which,

$$Av_i = \lambda_i v_i,$$

where v_i's are the (normalized, i.e. $||v_i||_2^2 = 1$ and orthogonal, i.e. $v_i^T v_j = \delta_{ij}$) eigenvectors of A, and λ_i are the corresponding eigenvalues.
Implication of $\lambda_i = 0 \implies \text{Rank Deficiency}$

The matrix A can be eigen-decomposed as,

$$A = Q \Lambda Q^{-1}$$

where, $Q = [\bar{v_1}, \bar{v_2}, \cdots, \bar{v_N}]$ and Λ is the matrix with diagonal entries as λ_i .

Tensors

Say, a tensor $T \in \mathbb{R}^{M \times N \times P}$ and a matrix $X \in \mathbb{R}^{M \times N}$.
For transpose of T, we also need to define the order of dimension swap, whereas in matrices, only 1 swap was possible (since, $2D$ matrices).

$$\text{e.g. } transpose(T, [0, 2, 1])$$

Functions

$$f : X \to Y$$

where, $x \in X$ and $f(x) \in Y$.

Continuity

If

$$\underset{\Delta x \to 0}{lt} f(x + \Delta x) = f(x)$$

and

$$\lim_{\Delta x \to 0} f(x - \Delta x) = f(x)$$

then, $f(x)$ is continuous at x.

Smoothness

If the derivative is continuous at x, then function is smooth at x.

Lipschitz Continuity

If

$$|f(x + \Delta x) - f(x)| \leq K \Delta x$$

for some $K \in \mathbb{R}$, then the function is Lipschitz continuous at x.

Derivative of a function

$$\frac{d}{dx} f(x)$$

Critical Points

The points where $f'(x) = 0$, which can lead to three types of critical points:

- Maxima - If $f''(x) < 0$
- Minima - If $f''(x) > 0$
- Inflection Point - If $f''(x) = 0$

Multi-Variate functions

$y = f(x_1, x_2)$, then gradient of the function is

$$\bigtriangledown f = \begin{bmatrix} \frac{df}{dx_1} \\ \frac{df}{dx_2} \end{bmatrix}$$

Thus, for calculating maxima and minima (or saddle point), we equate all entries of $\bigtriangledown f$ to 0.

Hessian Matrix

Now,

$$Hf = \begin{bmatrix} \frac{d^2f}{dx_1^2} & \frac{d^2f}{dx_1dx_2} \\ \frac{d^2f}{dx_2dx_1} & \frac{d^2f}{dx_2^2} \end{bmatrix}$$

All eigenvalues of Hf positive, then minima, if all negative, then maxima, else it is a saddle point.

Constrained Optimization using Lagrange Multiplier

If we want to maximize $f(x)$, then we will try to find the critical points and find the maxima. This would be known as unconstrained optimization.

For (equality) constrained optimization, maximize $f(x)$, subject to $g(x) = 0$.

Then, the direction of normals of $f(x)$ and $g(x)$ should be aligned in graphical representation.
Thus, we define a Lagrangian function

$$L(x) = f(x) + \lambda g(x), \lambda \neq 0$$

Making the gradient of L to be 0, we obtain λ.

$$\bigtriangledown L(x) = 0 \implies \bigtriangledown f(x) = -\lambda \bigtriangledown g(x)$$

where, substituting values of x obtained from $g(x) = 0$, will give the required λ.

Lecture 2

Name: Harsh Sanjay Roniyar

Roll Number: 22B3942

Random Variable

Example that we are considering is tossing a biased coin, hence
$\mu = P(X = 1) = 1 - P(X = 0)$ and they are not equal to $\frac{1}{2}$. The random variable here is X.

Probability Mass Function

Cumulative Distribution Function

Some common PMFs

1. Bernoulli Distribution

$x \in \{0, 1\}$
$\mu = P(X = 1)$

Hence, $Bern(x|\mu) = \mu^x(1 - \mu)^{1-x}$

2. Binomial Distribution

N coin tosses, with each toss with $P(X = 1) = \mu$

$Bin(y|\mu, N) = \binom{N}{y}\mu^y(1 - \mu)^{N-y}$

The event space $y \in \{0, 1, \ldots, N\}$
Then,

$$\bar{y} = \mathbb{E}(y) = \sum_{y=0}^{N} Bin(y|\mu, N) \cdot y$$

$$\mathbb{E}(f(y)) = \sum_{y} p(y)f(y)$$

Entropy

The entropy of y, would be defined as -

$$\mathbb{E}[-\log(p(y))]$$

which is equal to -

$$= -\mu \log \mu - (1-\mu)\log(1-\mu)$$

Joint, Conditional and Marginal Probability

- Joint: $p(y, z)$ such that $\sum_y \sum_z p(y, z) = 1$
- Conditional: $p(y|z) = \frac{p(y,z)}{p(z)}$
- Marginal: $p(z) = \sum_y p(y, z)$ and $p(y) = \sum_z p(y, z)$

Continuous Random Variable

For continuous random variable, $P(X = x) = 0$.

A pdf (probability density function) $p_X(x)$ describes the random variable.

Probability Density Function

- $p(x) \geq 0$
- $\int_x p(x)dx = 1$
- $p_X(x) = \frac{dP_X(x)}{dx}$

Cumulative Distribution Function

The CDF is describes as -

$$P_X(x) = \int_{-\infty}^{x} p_X(x)dx$$

The value of the CDF always reaches 1 at ∞ and starts from 0 at $-\infty$

Some common PDFs

1. Uniform Distribution

$$\mathrm{U}(x|a, b) - \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

2. Gaussian/Normal Distribution

$$\mathrm{N}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

3. Beta, etc.

Empirical Distribution

Empirical means *by experiment*.

Dirac-Delta function

$$\delta(x) = \lim_{\Delta x \to 0} \begin{cases} \frac{1}{\Delta x} & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

The empirical PDF is thus -

$$p_X(x) = \sum_{i=1}^{N} \frac{1}{N} \delta(x - x_i)$$

Integrating the PDF gives the CDF and similarly, differentiating CDF gives us the PDF.

Problems with Empirical Dist.

- Rote Learning
- No generalization

IID (independent and identically distributed)

- **Independent**: $p(x_i, x_j) = p(x_i)p(x_j)$
- **Identically**: $p_{X_i}(x_i = x) = p_{X_j}(x_j = x)$
  - $x_i \sim p_X$
  - $x_j \sim p_X$

Now, if we have $X = (x_1, x_2, \ldots, x_N)$
Then,

$$p(X) = p(x_1, x_2, \ldots, x_N)$$

$$= p_X(x_1)p_X(x_2)\ldots p_X(x_N)$$

$$= \prod_{i=1}^{N} p_X(x_i)$$

$$\log p(x) = \sum_i \log(p_X(x_i))$$

MLE (Maximum Likelihood Estimator)

For random variables, $x_1, x_2, \dots x_N$
ML Estimate of $\text{Gaussian}(\mu, \sigma)$

$$p(X) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Take log, then differentiate and equate to zero, to get μ_{ML} -

$$\mu_{ML} = \frac{1}{N} \sum_i x_i$$

For uniform distribution $U(x|a, b)$:

$$p(X) = \prod_i \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

On applying MLE, we get $a = \min x_i$ and $b = \max x_i$.

Lecture 3

Name: Harsh Sanjay Roniyar

Roll Number: 22B3942

Continuing from last class, discussing some examples for maximum likelihood estimation.

Example 1: Exponential Distribution

The ML estimate for the parameter λ turns out to be -

$$\lambda_{ML} = \frac{N}{\sum_i x_i} = \text{Inverse of the sample mean}$$

Sufficient Statistics

The smallest set of statistics that would give us the MLE of a parameter. Test Statistic is simply a function of the samples.

Some examples:

- Sample mean and variance for Gaussian distribution
- Sample mean for exponential distribution
- Max and min for uniform distribution

Non-parametric density estimation

Where a simple distribution doesn't fit the samples.
Can consider a mixture of multiple distributions.

$$p_X(x) = \frac{1}{N}\sum_{i=1}^{N} k(x - x_i),$$

where, $k(x) \geq 0, \forall x$, and $\int_{-\infty}^{\infty} k(x)dx = 1$
The function $k(\cdot)$ is called the kernel function. For all samples we consider their kernel functions and estimate density using the described method.

Assume a Gaussian Kernel

$$k(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}}$$

Select σ based on a rule of thumb that takes into account the given data (Avoid extreme cases).

1. MLE for parametric distribution

2. Bayesian estimation for parametric distribution

It takes into account a prior belief over parameters contrary to MLE. Many times, it is necessary to take into account previous beliefs instead of relying completely on the data.

Here, Assume a prior belief (distribution) $p_\Theta(\theta)$.
Now, instead of maximizing $\int \theta L_\Theta(X) d\theta$, maximize -

$$\frac{\int \theta p_\Theta(\theta) L_\Theta(X) d\theta}{\int p_\Theta(\theta) L_\Theta(X) d\theta}$$

The denominator term is there to keep the integral to 1.

Multivariate PDF

For a multivariable probability distribution, considering $2D$, $p(x_1, x_2) \geq 0$ and

$$\int \int p(x_1, x_2) dx_1 dx_2 = 1.$$

Also, marginal distribution over a reduced dimension would be -

$$p(x_2) = \int_{-\infty}^{\infty} p(x_1, x_2) dx_1$$

Conditional probability would be calculated as -

$$p(x_2 | x_1 = a) = \frac{p(x_1 = a, x_2)}{p(x_1 = a)}$$

where,

$$p(x_1 = a) = \int_{-\infty}^{\infty} p(x_1 = a, x_2) dx_2$$

Multivariate Gaussian

In this multi-dimensional representation (here, considering $2D$), the contours would be elliptical for the random variables covering the two axes. Also, in-place of variance, we would have a covariance matrix, which would help establish relationship between the two r.v.'s,

whether they are independent (correlated), or otherwise.

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22}^2 \end{bmatrix},$$

where $\sigma_{ij} = \sigma_{ji}$.
The σ_{ij}'s will be calculated as -

$$\sigma_{ij} = \mathbb{E}[(x_i - \mu_i)(x_j - \mu_j)]$$

The Gaussian distribution in this scenario has the following formula -

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k det(\Sigma)}} \exp\left(-\frac{(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)}{2}\right)$$

where, k is the data dimension, Σ is the covariance matrix and μ is the mean vector.

Exploratory Data Analysis

```
EDA is about taking stock of data
```

Things that we will check in data

- The entire data
- Each variable
- Pair of variables

Type of questions about each variable

- Type and coding
  - Nominal
  - Ordinal
  - True numerical
- Distribution
  - Descriptive statistics
  - Histograms
- Utility and ethics
  - Variability
  - Availability

Integers can be used to code nominal, categorical, ordinal, numerical and temporal data. In EDA, try to identify the description of discrete variables, such as the list of unique values, or order of values, etc.

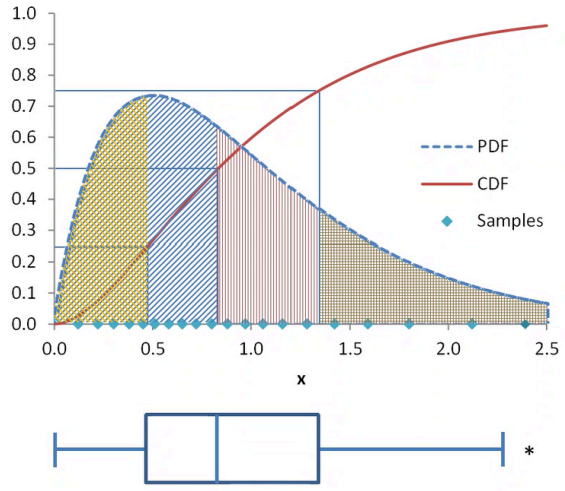
Data representations, such as histograms can help indicate anomalies, problems in the data.

Now, continuous variables are represented using PDFs, then the histogram divides the range into discrete **bins** for counting samples. It is similar to a stepwise approximation of a continuous distribution.

Mean is the Center-of-Gravity and Median divides the PDF into equal areas. Quartiles divides the PDF into four equal areas.

Right-Skewed (i.e. tail is longer on the right side of the peak) and *Left-Skewed* data.

Box and Whiskers plot summarizes the PDF. The three vertical lines in the box represent the 25th (***Q1***), 50th (***Q2*** `a.k.a` median) and 75th (***Q3***) percentile of the data. For the following image, it clearly shows that the data is right-skewed.



The Inter-Quartile Range (***IQR***) is defined as -

$$IQR = Q_3 - Q_1$$

> From above the upper quartile (***Q3***), a distance of 1.5 times the IQR is measured out and a whisker is drawn up to the largest observed data point from the dataset that falls within this distance. Similarly, a distance of 1.5 times the IQR is measured out below the lower quartile (***Q1***) and a whisker is drawn down to the lowest observed data point from the dataset that falls within this distance.

Box plot - Wikipedia

```
Correlation and scatter plots are between pairs of continuous variables
```

Correlation matrix can be computed for all variables together.

EDA using Python

```
import _libraries_
from _libary_ import __module__
```

Some important libraries for EDA (and ML in general)

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import scipy as sp
from sklearn import __module__
```

Lecture 4

Name: Harsh Sanjay Roniyar

Roll Number: 22B3942

Implementation of Data Science Techniques

- Understand the data using `pandas`
- Visualize data
- Comment Code
- Plot Histogram for the variables in the data
- Plot pair-wise scatter plots (Diagonals plot the histogram)
    - diagonals represent the histogram for the variables
    - off-diagonal terms represent the correlation between variables depending on the inclination angle.
- Plot the correlation matrix to see the correlation clearly
    - Can remove the variables having perfect correlation with an other variable.
- Plot side-by-side boxplot to identify variables which can help in our ML-related task such as classification. These variables can be termed as `discriminatory` variables

Basic Statistical Testing

Up until now, we were doing informal digging of data, known as exploratory data analysis. Now, we will test statistically for distributions to analyze data.

- IID assumption (done in Lecture 2)
- MLE of parametrized distribution (done previously in Lecture 3)
    - Examples for finding ML estimate

Now, statistical testing

Recipe for statistical testing

1. **Explore** - Explore reasonable assumptions about the data, e.g. distribution type (including “cannot be assumed”), mean, variance, etc. and ask what do we want to verify
2. **Null Hypothesis** - Form null hypothesis $\mathbf{H_0}$ that we want to reject, e.g. “The two means are NOT different”

3. **Alternative Hypothesis** - Form alternative hypothesis $\mathbf{H_1}$ that we hope is true, e.g. “The two means are different”
4. Decide on a significance level (1 confidence) to reject the null hypothesis BEFORE performing a test, e.g. p < 0.05 or p < 0.01
5. Perform the test by performing the calculations
6. Check if the result was significant enough to reject the null hypothesis and accept the alternative hypothesis, i.e., the alternative hypothesis was not just a chance outcome, but we are 95% or 99% confident that it is more likely than the null hypothesis

Confidence Interval

Given a sample $x_1, x_2, \ldots, x_N$ and sample mean $\bar{x}$

Find the interval $\bar{x} \pm \epsilon$ within which the true mean lies within confidence $1 - \alpha$ -

$$Pr(|\bar{x} - \mu| > \epsilon) < \alpha$$

When sample std. dev. is not known, replace with true std. deviation.

Comparing two independent set of samples

Welch's t-test:

$$t = \frac{\mu_X - \mu_Y}{\sqrt{\frac{\sigma_x^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$$

is matched to a table for the appropriate degrees of freedom (DoF):

$$\frac{\left(\frac{\sigma_x^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right)^2}{\frac{\sigma_x^4}{n_X^2(n_X-1)} + \frac{\sigma_Y^4}{n_Y^2(n_Y-1)}}$$

Comparing means of paired samples

Take the differences of the two sample values and store in a column Δ

With an assumed distribution

Without assuming a distribution

- Wilcoxon signed rank test :
  - Calculate Δ for the variables.
  - Add all ranks of positive and negative Δ separately (ranked in order of $|\Delta|$)

- Pick the smaller sum of ranks as test stat w_{test}.

$$w_{test} = \min\left(\sum_{i:d_i \geq 0} r_i, \sum_{j:d_j < 0} r_j\right)$$

- Test stat w_{test} should be smaller than $w_{critical}$ (obtained from the table) for the given N
- Compare the Δ distribution.

Linearly related variables

Two paired continuous variables

```
Correlation does not imply Causation
```

Pearson's Correlation Coefficient:

- Ranges from -1 (perfectly negative correlated) to +1 (perfectly positive correlated)
- Represented by $\rho_{X,Y}$

$$\rho_{X,Y} = \frac{\operatorname{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[X,Y] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\mathbb{E}[X^2] - \mathbb{E}[X]^2}\sqrt{\mathbb{E}[Y^2] - \mathbb{E}[Y]^2}}$$

- For a sample

$$r_{x,y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

  Here, we will be using **Spearman's Correlation Coefficient** which is same as Pearson's correlation coefficient but for **ranks**. Used for non-parametric testing.

$$r_S = \rho_{R(X),R(Y)} = \frac{\operatorname{Cov}(R(X),R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}} = 1 - \frac{6\sum(R(x_i) - R(y_i))^2}{n(n^2-1)}$$

Choosing a stat test

- Frame your problem
  - Predictor and Outcome variable
- Check for widely acceptable tests among data-scientists.
- Check if the assumptions made by the test hold for your scenario.
- Else, make your own test by using an existing test as a base.

Goal: Check for statistical significance

Graphs

Basic Graphs

- Bar Graph
- Line Chart
- Pie Chart
- Scatter Plot

Chart Information

- Chart Title
- Axes Titles
- Axes Units
- Grid Lines (Horizontal and/or Vertical)
- Legend
- Error Bars
- Confidence Intervals

Graph Beautification

- Font Legibility
- Font Consistency
- Information-to-Ink Ratio (Add whitespace)
- Color Usage
- Sorting (Value/Alphabetical)
- Highlighting
- Boxes (to highlight specific regions in graphs)
- Call-Outs and Pointers
- Breakout Pie-Chart

Advanced Graphs

- Histogram
- Stacked Bar (Value or Percentage)
- Column Chart
- Area Chart
- CDF
- Dual Y-Axis Bar/Line Charts

- Donut Charts
- Mixed Charts
- Bubble Chart (3-D curves)
- Kaplan-Meier (or Survival) Curves - to show when failure occurs
- Box and Whisker Plots
- Violin Plot - shows several distributions together
- Gantt Chart
- Radar or Spider Chart, etc.

What to use?

- Dimensionality Reduction for high-dimensional data
- Data Transformations (log-scale, ratio, polar form)

HOMEWORK: RDBMS and SQL

Lecture 5

Name: Harsh Sanjay Roniyar

Roll Number: 22B3942

Q-Q Plot (Quantile-Quantile Plot)

Quantile ranges from 0 to 1 ($\sim$ to percentiles).
The plot graphs data quantiles (x) vs theoretical quantiles (z) .

Typically, used for normal distribution, can be extended to other distributions also by changing the z-distribution on the horizontal axis.
If the data is Gaussian-distributed, we will get a straight line on the Q-Q plot.

If the graph cuts the straight line from the x-side, then it is left skewed, else right skewed. This can be seen with the help of the distinguishing compression for $< 50\%ile$ and above them.

Database! What is a Database?

RDBMS (Relational Database Management System)

- Tables
- Records
- Fields
- Keys

SQL (Structured Query Language)

Standard language for interacting with (R)DBMS

- CREATE TABLE
- INSERT INTO
- DELETE FROM

Typical query structure in SQL:

- FROM table
- WHERE condition
- SELECT

Intro To DS Completed

ML for Smart Monkeys

ML acts as an estimator by using data to create models that can make new predictions about similar data.

```
Sweet Spot for ML: Lots of Structured Stationary Data
```

Parameters and Hyper-Parameters

- Parameters: variables whose values are updated during training
- Hyperparameters: whose values are fixed by model developer before beginning of learning process.

Types of ML Problems:

- Supervised Learning
- Unsupervised Learning
- Semi-Supervised Learning, Reinforcement Learning, etc.

ML Recipe

- Type of ML problem
- Data Preparation
- ML Framework Selection
- Prepare training, test, validation data
- Perform training, validation and testing

Bias-Variance Trade-Off

- **Underfitting**: High bias, Low variance
- **Overfitting**: Low bias, High variance

Lecture 6

```
Assignment 1: Released
Deadline: 1st September 2024
```

Name: Harsh Sanjay Roniyar

Roll Number: 22B3942

Regularization

Penalizing overfitting to the training data curve, in some sense constraining the model to reduce the complexity of the model.

Loss Function on including the **L2 Norm** for regularization becomes -

$$L = L_{\text{old}} + \lambda \sum_{i=1}^{N} w_i^2$$

Note that regularization hurts training set performance! This is because it limits the ability of the network to overfit to the training set. But since it ultimately gives better test accuracy, it is helping your system

Performance Metric

A judge for the model

Data Preparation

1. **Train** - Learn (optimize parameters)
2. **Validation** - Compare hyperparameters
3. **Test** - One final model evaluation

Having more data-points reduce chances of overfitting and thus less variance.

K-Fold Cross Validation

1. Train K times on K-1 folds and each time validate on the hold-out fold.
2. Select hyperparameters

3. Train on all K folds using the selected hyperparameters (can use Train + Validation data together).
   Thus, validation is done and now final test can be done using the test data.

Lecture 7

Name: Harsh Sanjay Roniyar

Roll Number: 22B3942

Linear Regression

Data: $[x_1 \dots x_D]$, t
Predictions: $y = f_w(\bar{x})$
where, $t_i, y_i \in \mathbb{R}, \forall i$

$$y_i = \left(\sum_{j=1}^{D} w_j x_{ij} \right) + w_0 = W^T x$$

This acts as a linear approximation of our data.

Since $x_i \in \mathbb{R}$. Then I can create $D+1$ features in $\phi_i = \begin{bmatrix} 1 \\ x_i \\ x_i^2 \\ \vdots \\ x_i^D \end{bmatrix}$. Then this is a polynomial representation in the x-space, but in linear form in the space ϕ.

For (all) the data-points, fitting a Gaussian curve at those points and then adding the distributions also gives us a fit for the data.

$$\phi = \left[e^{-\frac{(x-x_j)^2}{2\sigma^2}} \right]$$

where ϕ is a $W \times N$ matrix which can be reduced to $W \times D$.

Radial Basis Function. Similar to Kernel Density Estimation.

Now, back to linear regression.
The predictions t, would have some stochastic noise.

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon$$

Assuming the noise term to be Gaussian and $\beta = \frac{1}{\sigma^2}$

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{i=1}^{N} N(t_i|\mathbf{w}^T\phi(x_i), \beta^{-1})$$

Now, since this has product terms, taking $\log$:

$$\ln p(\mathbf{t}|\mathbf{w}, \beta) = \sum_{i=1}^{N} \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{\{t_i - \mathbf{w}^T\phi(x_i)\}^2}{2\sigma^2}$$

$$E_D(w) = \frac{1}{2}\sum_{i=1}^{N}\{t_i - \mathbf{w}^T\phi(x_i)\}^2$$

Now, going on to maximizing the likelihood - for this we need to set the gradients w.r.t x to zero. So, we are maximizing,

$$\nabla \ln p(\mathbf{t}|\mathbf{w}, \beta) = \sum_{i=1}^{N}\{t_n - \mathbf{w}^T\phi(x_n)\}\phi(x_n)^T$$

which finally gives,

$$\mathbf{w_{ML}} = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{t}$$

where,

$$\Phi = \begin{bmatrix} \phi(x_1)^T \\ \phi(x_2)^T \\ \vdots \\ \phi(x_N)^T \end{bmatrix}$$

Conclusions:

1. Assume linear model
2. Assume error is Gaussian distributed
    1. M.S.E is the metric to maximize data likelihood
    2. The analytical solution for $w = \text{PseudoInv}(\mathbf{w}) \times \mathbf{t}$

Bias-Variance Decomposition

`of linear regression and gaussian noise`

$$L_i = \text{Loss} = (y(x_i) - t_i)^2$$

Then,

$$\mathbb{E}(L) = \int\int L \cdot p(x,t)dxdt = \int\int (y(x) - t)^2 p(x,t)dxdt$$

$$\frac{\delta \mathbb{E}(L)}{\delta y(x)} = 2 \int \{y(x) - t\} p(x,t) dt = 0$$

This gives,

$$y(x) = \frac{\int t p(x,t) \mathrm{d}t}{p(x)} = \mathbb{E}_t(t|x)$$

Now, if we try to do manipulation with

$$\{y(x) - t\}^2 = \{y(x) - \mathbb{E}(t|x) + \mathbb{E}(t|x) - t\}^2$$

$$= \{y(x) - \mathbb{E}(t|x)\}^2 + \{\mathbb{E}(t|x) - t\}^2 + 2\{y(x) - \mathbb{E}(t|x)\}\{\mathbb{E}(t|x) - t\}$$

Lecture 8

Name: Harsh Sanjay Roniyar

Roll Number: 22B3942

Continuing from where we left in the last lecture...

The nature of the bias-variance relations can be better visualized using the contour plots of $y, t, \text{and } x$.

Now, the calculation of the expected loss, as left in Lecture 7 becomes -

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[t|x]\}^2 p(x) dx + \int \{\mathbb{E}[t|x] - t\}^2 p(x) dx$$

The second term in this loss, can be related to the irreducible noise of our data. We can attempt to reduce the first part of the expected loss from our end.

$$h(x) = \mathbb{E}[t|x] = \int t p(t|x) dt$$

Now, $\{y(x) - h(x)\}$ is meaningless since our $y(x)$ depends on the data samples. Hence, we will represent it using $\{y(x;D) - h(x)\}^2$ where D is the sample of the distribution.

$$= \{y(x;D) - \mathbb{E}_D[y(x;D)] + \mathbb{E}_D[y(x;D)] - h(x)\}^2,$$

where $\mathbb{E}_D[y(x;D)]$ is the expectation over the model. $h(x)$ can be seen as the best theoretical model, and $y(x;D)$ is the particular model under consideration for the given training data.

On further simplification, the final expression reduces to -

$$\mathbb{E}_D\left[\{y(x;D) - h(x)\}^2\right] = \underbrace{\{\mathbb{E}_D[y(x;D)] - h(x)\}^2}_{\texttt{(bias)}^2} + \underbrace{\mathbb{E}_D[\{y(x;D) - \mathbb{E}_D[y(x;D)]\}^2]}_{\texttt{variance}}$$

Hence, the expected loss can be written as $\texttt{(bias)}^2 + \texttt{ variance } + \texttt{ noise}$. The noise relation will be

$$\texttt{noise} = \int \{h(x) - t\}^2 p(x,t) \mathrm{d}x \mathrm{d}t$$

Model Regularization

L2 Regularization

$$y(x_i) = \sum_{j=1}^{D} w_j x_i j + w_0$$

Then, the L_p norm of $\mathbf{w}$ is given by

$$\left(\sum_{j=1}^{D} |w_j|^p \right)^{\frac{1}{p}}$$

The norms of L_∞ = max $|w_j|$ and L_0 = count of non-zero components.

The objective function is -

$$\min_{w} \frac{1}{2} \sum_{i=1}^{N} (y(x_i) - t_i)^2 + \frac{\lambda}{2} \sum_{j=1}^{D} |w_j|^q$$

with,

$$\sum_{j=1}^{D} |w_j|^q \leq \eta$$

The objective can be further written as -

$$\min_{\mathbf{w}} \frac{1}{2} \sum_{i=1}^{N} (t_i - \mathbf{w}^T \phi(\mathbf{x}_n))^2 + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w}$$

$$\implies E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}).$$

The optimized solution for $\mathbf{w}$ is -

$$\boxed{\mathbf{w} = \left(\lambda \mathbf{I} + \Phi^T \Phi\right)^{-1} \Phi^T \mathbf{t}}.$$

`The dimension of` Φ `is` $N \times D$. `Hence, we are taking the pseudo-inverse.`

Lecture 9

Name: Harsh Sanjay Roniyar

Roll Number: 22B3942

Convex Functions

Convex Functions are defined mathematically, as follows-

$$\alpha L(w_1) + (1-\alpha)L(w_2) \geq L(\alpha w_1 + (1-\alpha)w_2),$$

where, α is between 0 and 1.
In L2 regularization, we work under the assumption that $L(w)$ is convex.

Adding two convex functions, with a positive multiplier for each, the resulting function is also a convex function. (`Jensen's Inequality`).

Gradient Descent Algorithm

A very important technique in all of modern AI-ML tasks.

Basic intuition - Even if we can't calculate the exact minimum, we can still get closer to the minimum iteratively in small steps.

Randomly initialize w
Until we reach terminal condition.

$w_{\text{new}} \leftarrow w_{\text{old}} - \eta \nabla L_w$
$w_{\text{old}} \leftarrow w_{\text{new}}$

η is the learning rate (or step size) and needs to be tuned properly in real-world systems.

Perform this in all components of w (dimension of w).

Now, the terminal conditions could be one (or a combination) of the following -

1. `max_iter` has been reached
2. $L(w_o) - L(w_n) < \epsilon$
3. $(w_o - w_n)^T (w_o - w_n) < \delta$

Since, from the last lecture Lecture 8, we has the expressions for E_D and E_W, hence we can compute the gradient for those expressions and we would get the update-iterate step.

Lecture 11

Name: Harsh Sanjay Roniyar

Roll Number: 22B3942

$x^T M x + x^T v + C$

Linear: $x^T v + c = 0$

Let's come back to 1 dimension..

The decision boundary is $W^T x + b$ and depending on the sign of the boundary for a specific value, the data-point is assigned to that class.

Suppose, we did have two classes in the single dimension, but the σ for these classes is very different. Their intersections give some sort of thresholds (boundaries) for the classifiers. In this case, we won't be able to get a linear classifier since we have two threshold conditions.

Thus, even if we have different σ's in one of the dimensions, the whole classifier for the n-dimensional data cannot retain its non-linearity.

Now, in the case when we were in 2 dimensions, and have unequal σ's, then the bayesian boundaries, wouldn't be exactly same contours. In order to find the boundaries, we assume the distributions to be linear.

Gradient Descent for Linear Classifiers

For Regression, we had loss options as - 1. MSE and 2. MAE and L2 and L1 regularization.

Now, for linear classifiers, we have the following Loss function -

1. Misclassification Rate:

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{sign(y_i) \neq t_i\}$$

   where $\mathbb{1}$ is the indicator function, $t_i \in \{-1, 1\}$ and $y_i = w^T x_i + b$. The indicator function is a binary function, where it outputs 1 if the condition evaluates to true, else 0. But, the problem with the sign function is that its is almost zero everywhere, leading to a discontinuous classifier.

2. Replacing $\text{sign}(y_i)$ with $\sigma(y_i)$ where σ is the sigmoid function given by -

$$\sigma(y_i) = \frac{1}{1 + e^{-y_i}},$$

and we also change the labels t_i from $\{-1, 1\}$ to $\{0, 1\}$.
Now, $y_i = \sigma(w^T x_i + b)$ interpreted as $p(t = 1|x_i)$ and thus, $y_i \in (0, 1)$ and $w^T x_i + b \in (-\infty, \infty)$.

3. Measuring loss as

$$\text{loss} = \frac{1}{2N} \sum_i (y_i - t_i)^2$$

can work but its error is gaussian, and might not give the most optimal result.

4. **KL Divergence** - BCE (Binary Cross Entropy):

$$-\sum_i \sum_c t_{c,i} \log y_{c,i} = -\sum_i [t_i \log y_i + (1 - t_i) \log(1 - y_i)]$$

where, c is the class.

Logistic Regression - Minimize BCE of sigmoid of line lin. exp. wrt. a binary target.

Gradient Descent using BCE

$$\frac{dl_i}{dw} = \frac{dl_i}{dy_i} \frac{dy_i}{dh_i} \frac{dh_i}{dw}$$

$$= \left(\frac{t_i}{y_i} - \frac{1 - t_i}{1 - y_i} \right) \cdot y_i \cdot (1 - y_i) \cdot x_{i,k}$$

Lecture 12

Name: Harsh Sanjay Roniyar

Roll Number: 22B3942

Continuing from where we left in Lecture 11,

$$\frac{dL}{dw} = \frac{dL}{dy}\frac{dy}{dh}\frac{dh}{dw};$$

where, $h = wx$, $y = \sigma(h) = \frac{1}{1+e^{-h}}$ and $L = -t\log(y) - (1-t)\log(1-y)$
Therefore, $\frac{dL}{dy} = \left(-\frac{t}{y} + \frac{1-t}{1-y}\right)$, $\frac{dy}{dh} = y(1-y)$.
Hence, we get

$$\frac{dL}{dw} = (y-t)x.$$

In linear as well as logistic regression, the model is still linear and loss functions are also convex.

Now, considering regularization, for L2 Regularization, the total loss comes out to be $L_{\text{error}} + \frac{\lambda}{2}w^T w$, whereas in L1 regularization, it is $L_{\text{error}} + \lambda\sum_j |w_j|$.
Also, L2 regularization has some sort of weight stabilization effect, i.e. it tries to correct the deviations of w_i's such that the loss is minimized, whereas L1 has no such effect.

Elastic Net

`Best of both L1 and L2 (but there's obviously a catch)`

$$\text{Total Loss } = L_{\text{error}} + \frac{\lambda_2}{2}\sum_j |w_j|^2 + \lambda_1 \sum_j |w_j|$$

The catch here is - Now we need to tune two hyperparameters instead of just one.

Confusion Matrix

Predicted \ Actual	0	1
0	TN	FN
1	FP	TP

Symmetric Risk

$(0,0), (1,1)$ - Low Risk

Predicted \ Actual	0	1
0	low	high
1	high	low

Asymmetric Risk

Different risks for **Type I (FP)** and **Type II (FN)** errors.

Predicted \ Actual	0	1
0	low	higher
1	high	low

Some metrics for binary classification

Single Threshold

$$\text{Accuracy} = \frac{TP + TN}{\text{Total}}$$

$$\text{Precision (P)} = \frac{TP}{TP + FP}$$

$$\text{Recall Sensitivity (R)} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Balanced Metric : } F1 \text{ score} = \frac{2PR}{P + R}$$

All the metrics defined above are threshold dependent, depending on the boundary condition set for classification between the classes 0 and 1, the values for all metrics will change accordingly.

Threshold-free Metric

AUC (Area Under (receiver operating characteristic) Curve)

The plot is made between sensitivity and $(1 - \text{specificity})$. The farther the threshold point from the origin, the lower the threshold (variation of both values on the axes from 0 to 1).

```
Accuracy is meaningful only when we have balanced classes otherwise
preferable to use other metrics like AUC and F1-score.
```

Jensen's Inequality

Similar to what we did in Lecture 9.

A function f is convex iff $\forall\ 0 \leq \lambda \leq 1$ -

$$\lambda f(x_1) + (1-\lambda) f(x_2) \geq f(\lambda x_1 + (1-\lambda) x_2)$$