Multiple Regression

Given

$\mathcal{D} = \{ x_i, y_i \}_{i=1}^{n}$

$x_i \in \mathbb{R}^d$ Multiple independant variable

$y_i \in \mathbb{R}$ Single dependant variable

$y_i = f(x_i)$

Objective :

$\hat{f} : \mathbb{R}^d \to \mathbb{R}$

Hypothesis

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_d x_d + \varepsilon$ Random error

$\beta_i \Rightarrow$ slope of respective axis

$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_d x_d + e$ error in the model

(data points) $i = 1$ to n

$j = 1$ to d (feature)

$Y = Xw + e$

$$w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix}_{(d+1)\times 1} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{1 \times n}$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \dots & x_{1d} \\ 1 & x_{21} & x_{22} & x_{23} & \dots & x_{2d} \\ \vdots & & & & & \\ 1 & x_{n1} & x_{n2} & x_{n3} & \dots & x_{nd} \end{bmatrix}_{n \times (d+1)}$$

Find w

$$SSE = \min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\min SSE = \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_{i1} + w_2 x_{i2} + \dots + w_d x_{id}))^2$$

UCOP

$$\frac{\partial L}{\partial w_0} = \sum_{i=1}^{n} 2(y_i - (w_0 + w_1 x_{i1} + w_2 x_{i2} \ldots w_d x_{id}))(-1) = 0$$

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (w_0 + w_1 x_{i1} + w_2 x_{i2} \ldots w_d x_{id})$$

$$\sum y_i = n w_0 + w_1 \sum_{i=1}^{n} x_{i1} + w_2 \sum_{i=1}^{n} x_{i2} \ldots w_d \sum_{i=1}^{n} x_{id}$$

$$\frac{\partial L}{\partial w_1} = \sum_{i=1}^{n} 2(y_i - (w_0 + w_1 x_{i1} + w_2 x_{i2} + \ldots w_d x_{id}))(-x_{i1}) = 0$$

$$\sum y_i x_{i1} = w_0 \sum x_{i1} + w_1 \sum x_{i1}^2 + w_2 \sum x_{i1} x_{i2} + \ldots w_d \sum x_{i1} x_{id}$$

$$\frac{\partial L}{\partial w_2} = \sum_{i=1}^{n} 2(y_i - (w_0 + w_1 x_{i1} + w_2 x_{i2} + \ldots w_d x_{id}))(-x_{i2}) = 0$$

$$\sum y_i x_{i2} = w_0 \sum x_{i2} + w_1 \sum x_{i1} x_{i2} + w_2 \sum x_{i2}^2 + \ldots w_d \sum x_{id}^2$$

$$\frac{\partial L}{\partial w_d} = \sum_{i=1}^{n} 2(y_i - (w_0 + w_1 x_{i1} + w_2 x_{i2} + \ldots + w_d x_{id})(-x_{id}) = 0$$

$$\sum y_i x_{id} = w_0 \sum x_{id} + w_1 \sum x_{i1} x_{id} + \ldots \ldots w_d * \sum x_{id}^2$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & \ldots & 1 \\ x_{11} & x_{21} & \ldots & \ldots & x_{n1} \\ x_{12} & x_{22} & \ldots & \ldots & x_{n2} \\ \vdots & \vdots & & & \vdots \\ x_{1d} & x_{2d} & \ldots & \ldots & x_{nd} \end{bmatrix}_{(d+1) \times n}$$

$$X^T X = \begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \ldots & \sum_{i=1}^{n} x_{id} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1} x_{i2} & \ldots & \sum_{i=1}^{n} x_{i1} x_{id} \\ \sum x_{i1}^2 & \sum x_{i1} x_{i2} & \sum x_{i2}^2 & \ldots & \sum x_{i2} x_{id} \\ \vdots & & & & \\ \sum x_{id} & \sum x_{id} x_{i1} & \sum x_{id} x_{i2} & & \sum x_{id}^2 \end{bmatrix}$$

$$X^T Y = \begin{bmatrix} \sum y_i \\ \sum x_{i1} y_i \\ \sum x_{i2} y_i \\ \vdots \\ \sum x_{id} y_i \end{bmatrix}_{(d+1) \times 1}$$

$$X^T X w = X^T Y$$

$$w = (X^T X)^{-1} X^T Y$$

$$w = (X^T X)^{-1} X^T Y$$

Cost	7	3	3	4	6	7
Distance	560	220	340	80	150	330
Time	16.68	11.50	12.03	14.88	13.75	18.11

$$X = \begin{bmatrix} 1 & 7 & 560 \\ 1 & 3 & 220 \\ 1 & 3 & 340 \\ 1 & 4 & 80 \\ 1 & 6 & 150 \\ 1 & 7 & 330 \end{bmatrix} \qquad X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 7 & 3 & 3 & 4 & 6 & 7 \\ 560 & 220 & 340 & 80 & 150 & 330 \end{bmatrix}$$

$$Y = \begin{bmatrix} 16.68 \\ 11.50 \\ 12.03 \\ 14.88 \\ 13.75 \\ 18.11 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 6 & 30 & 1680 \\ 30 & 1680 & 9130 \\ 1680 & 9130 & 615400 \end{bmatrix} \qquad (X^T Y)^{-1} = \begin{bmatrix} 1.60723 & -0.25 & -6 \times 10^{-4} \\ -0.25 & 0.0698 & -3 \times 10^{-4} \\ -6 \times 10^{-4} & -3 \times 10^{-4} & 8.6 \times 10^{-6} \end{bmatrix}$$

$$X^T Y = \begin{bmatrix} 86.95 \\ 456.14 \\ 25190.2 \end{bmatrix} \qquad w = (X^T X)^{-1} X^T Y$$

$$w = \begin{bmatrix} 8.5655 \\ 1.1965 \\ -2 \times 10^{-4} \end{bmatrix}$$

$$\hat{y} = 8.5655 + 1.1965 x_1 + (-2 \times 10^{-4}) x_2 + e.$$

Assumptions

* No multicolinearity (No correlative variables)

Measures

SSE

SST

SSR

R^2

Adjusted $R^2 = 1 - \frac{(1-R^2)(n-1)}{(n-d-1)}$ Increase d Adjusted R^2 ↓se

$= 1 - \frac{SSE/n-(d+1)}{SST/n-1}$

i> Correlations

Check for each pair of features

$-1 \le \underset{\text{Correlation}}{\sigma_{AB}} = \frac{\overset{\text{Covariance}}{\Sigma_{AB}}}{\underset{\text{Standard Deviation}}{\sigma_A \ \sigma_B}} \le +1$

+1 = Positive correlated (A ↑se then B ↑se)

0 = Independent.

Construct Correlation matrix

$\begin{bmatrix} & \\ & \end{bmatrix}_{d \times d}$ [d = features]

$-0.7 \le \sigma_{ij} \le 0.7$

ii> Variance Inflation factor

$VIF = \frac{1}{1-R^2}$

$\underbrace{x_1, x_2, x_3, x_4 \ldots x_d}$ Y

$x_i \ x_j$ ⇒ keeping one as independent & other dependent
Construct linear reg model & find VIF

No of pairs = dC_2
(No of models to construct)

$\begin{bmatrix} & \\ & \end{bmatrix}_{d \times d}$

Highly correlation VIF is high ∴ remove 1 feature

HP in polynomial reg
→ degree of polynomial
→ ?.

$d=1$ $d=2$ ... $d=10$

↓

Training

↓

M_1 M_2 ... M_{10}

↓

Validation — choose optimal value of HP

↓

R_1 R_2 ... R_{10} $M^* = \arg\max_{k} \{R_k^2\}$ $1 \le k \le K$

(High R^2 - best)

$M^* = d^*$

↓

Testing

↓

R^2 - actual value

Error

- Training Error
- Test error / True error / Generalised error

Traing Error $\begin{cases} w \\ \text{min SSE} \end{cases}$

TE = 0 ; Fn passes through all data points

TE = Large : Doesnot pass through many data points

Test error ≥ 0 ; Test data points lie on Line.

Bias = Training error

Variance = Test error (How varied from Training error)
(deviated)

Bias

- Low = model fits train data well / Complex model (Polynomial degree high)
- High = model doesnot fit train data well

Variance

- Low = less deviation from Training error / Good model

Var \ Bias	Low	High
Low	Ideal	×
High	Complex Over fitting	Under fitting

Finding the optimum value of H.P. (Hyper parameter) – 3 way split

Model Selection
- Hold out
- k fold Cross validation
- Leave - one - out
- 3 way split

DS → Construct: Train → [ML] → model

DS → Evaluate: Test → model

Hold out → split 2 : 1

Size (Train DS) > Size (Test)

Experiment k times & x^2 avg taken

k fold Cross validation

Split in k folds

At i experiment ith is used as Test set other fold used as Train set

Train set and Test set at any point must be disjoint

| F_1 | F_2 | . . | . . | F_k |

Every fold tested & remain used as Train set

Lastly find avg

Leave one out (n Fold cross validation)

n folds
Each fold consist of 1 data point

Every point tested while other n-1 points kept as Trainning

n is large (Data set large) – Hold out, k fold

Data set small – Leave one out

3 way split

Split in 3 data sets – Training, Validation, Test

Reasons for overfit & Under fit

1) Insufficient Trainset
2) Noise / outlier

Increase data set

- Augmentation - Do transformation
- Smote analysis

Overfitting (overcome)

- Ridge Regression / L_2 - Regularisation
- Lasso " / L_1 - "
- Elastic net

→ To all parametric mode

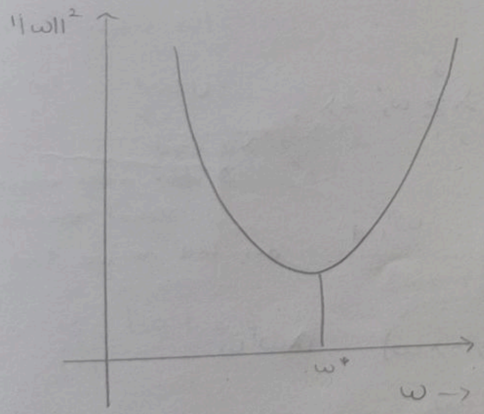
Find

$$\underset{\omega}{\min} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \text{Regularization term}$$

$$+ \ \|\omega\|^2$$

$$\omega^T = (\omega_0 \dots \omega_d)$$

$$\|\omega\| = \sqrt{\omega_0^2 + \omega_1^2 + \dots + \omega_d^2} \quad - L_2 \text{ Norm}$$



Ridge regularisation

Find ω

$$\min \quad SSE + \lambda \cdot L_2 \text{ norm of } \omega$$

⇓

Find ω

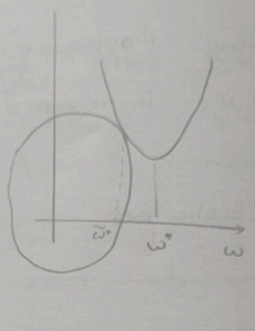
$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i) + \lambda \|\omega\|^2$$ → square for easier derivation

Find ω
$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{d} \omega_j^2$$

Loss function : Continuous & Differentiable fn

Ridge Regularisation   D-dimensional sphere

Prerequisite (Not consider bias)
Data points to be normalised
And the center



$\frac{X-\mu}{\sigma}$

Find ω
$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$\Downarrow$

Find ω
$$\min \sum_{i=1}^{n} (y_i - \omega^T x_i)^2 \quad \text{Convex fn}$$

s.t.c
$$\omega_1^2 + \omega_2^2 + \omega_3^2 + \dots + \omega_d^2 = 1 \quad \text{Convex set}$$

} Constrained Convex optimisation problem

$\Downarrow$

Legrang multiplier ⇒ used to convert Constrained OP to unconstrained OP

Fin ω
$$\min \sum_{i=1}^{n} (y_i - \omega^T x_i)^2 + \lambda \left(\sum_{j=1}^{d} \omega_j^2 - 1 \right)$$

$\Downarrow$

$$\sum_{i=1}^{n} (y_i - \omega^T x)^2 + \lambda \sum_{j=1}^{d} \omega_j^2 - \lambda \qquad \text{UCOP}$$

λ is const
It won't affect value of obj fn
Hence removed.

$$\boxed{Y_{n\times 1} \quad X_{n\times d} \quad \omega_{d\times 1} \quad \omega^T_{1\times d}}$$

d+1 tod as we donot include 1

Find ω

$$L = \overbrace{(Y - X\omega)^T (Y - X\omega)}^{(Y-X\omega)^2} + \lambda \omega^T \omega$$

$n\times 1$ $n\times d$ $d\times 1$ $1\times n$ $n\times 1$

Matrix Notation
$X^2 = X^T X$

$(Y^T - \omega^T X^T)(Y - X\omega)$

$$L = (Y^T Y) - \underline{Y^T \cancel{\omega X} X\omega} - \underline{\cancel{\omega} X^T Y} + \omega^T X^T X \omega + \lambda \omega^T \omega$$

Equal (Transpose of one another)

$$= \cancel{X^T Y} - \cancel{2 X \omega^T Y} + \cancel{X^T \omega \omega X} + \cancel{\lambda \omega^T} \omega$$

$$= Y^T Y - 2\omega^T X^T Y + \omega^T X^T X \omega + \lambda \omega^T \omega$$

$$= Y^T Y - 2\omega^T X^T Y + \omega^2 X^T X + \lambda \omega^2$$

$$\frac{\partial L}{\partial w} = -2X^TY + 2wX^TX + 2\lambda w$$

$$= -X^TY + \cancel{w}(X^TX + \lambda I)w = 0$$

$$\cancel{w}(X^TX + \lambda I)w = X^TY$$

$$w = (X^TX + \lambda I)^{-1}X^TY$$

$\lambda = 0$ Similar to Multiple Regression.

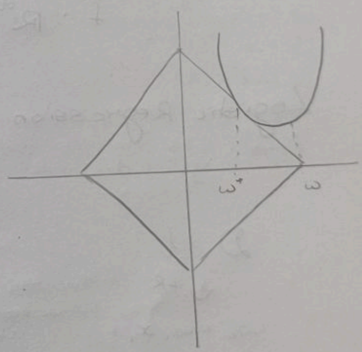
Overfitting not solved.

w_j explains the importance of feature x_{ij}

Ridge regularisation ⇒ Solution to Multicolinearity
Ridge regularisation } ?

∴ the importance of features ↓

Lasso regularisation. (feature selection technique)



Min

SSE + L_1 Norm

$$\sum(y_i - \hat{y}_i)^2 + \lambda\left(\sum_{j=1}^{d}(|w_j|)\right)$$

Obj = Convex fn

Constraints = Convex set.

∴ COP

⇓

~~Stc.~~

$$\text{Min} \sum_{i=1}^{n}(y_i - w^Tx_i)^2$$

Stc. $\sum_{j=1}^{d}|w_j| = 1$

⇓

$$\text{Min} \sum_{i=1}^{n}(y_i - w^Tx_i)^2 + \lambda\left(\sum_{j=1}^{d}|w_j| - 1\right)$$

⇓

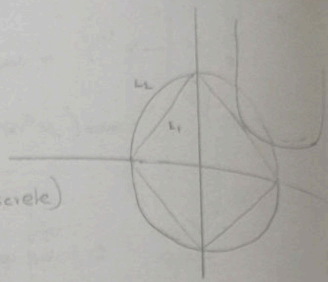
$$\text{Min} \sum(y_i - w^Tx_i)^2 + \lambda\sum_{j=1}^{d}|w_j|$$

// $w = 0$ feature not selected ∴ used to Select feature

$$w = (X^TX + \lambda I)^{-1} \cdot X^TY$$

Elastic Net

Both L_1 & L_2 norm added.



Classification (class variable - discrete)

Given

$$\mathcal{D} = \{x_i, y_i\}_{i=1}^{n}$$

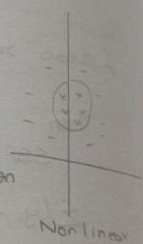
$$x_i \in R^d \text{ [Discrete / Continuous]}$$

$$y_i \in \{c_1, c_2\}$$

$$f : x_i \rightarrow y_i$$

Objective

$$\hat{f} : \mathbb{R}^d \rightarrow \{c_1, c_2\}$$



Logistic Regression - attempt to apply regression in classification

y_i = discrete

$\mathcal{D} = \{$

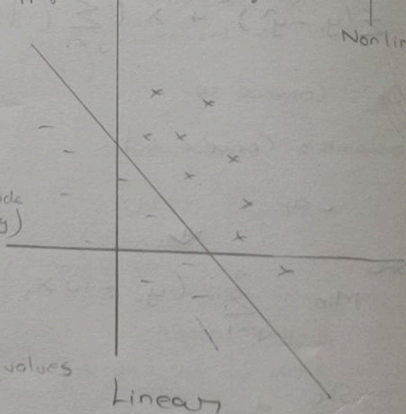
↓ LR (Min SSE but we need line to divide exactly)

$w_0 + w_1 x_1 + \dots \; w_d x_d$

↓

× $y = c_1 / c_2$

$\mathbb{R}$ but gives continuous values



to ↓ $w^T x$ change to discrete

Sigmoid function

$0 < \sigma(x) < 1$ for $-\infty \leq x < +\infty$

(Range of 0 to 1)

↓ threshold value 0.5

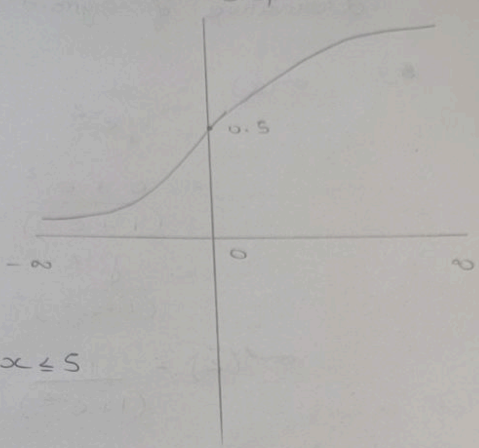
$$\hat{y} = \begin{cases} c_1 & \text{if } \sigma(w^T x) \geq 0.5 \\ c_2 & \text{o/w} \end{cases}$$

Sigmoid function

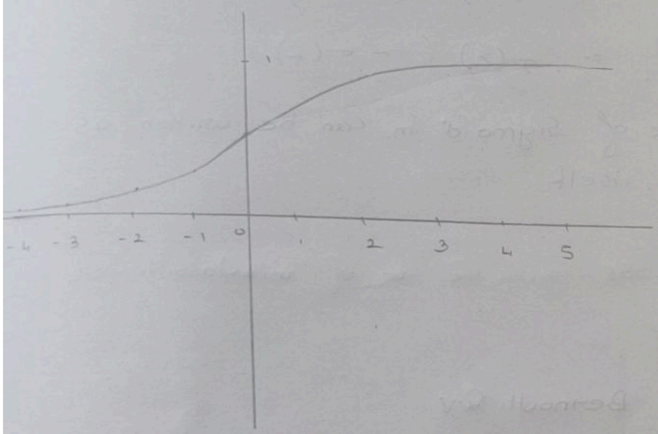
$$\sigma(z) = \frac{1}{1+e^{z}}$$

$$\sigma(0) = 0.5$$

S shaped fn



Plot for $\sigma(x)$: $-5 \le x \le 5$



$\sigma(-1) = 0.26$

$\sigma(-2) = 0.11$

$\sigma(-3) = 0.04$

$\sigma(-4) = 0.01$

$\sigma(-5) = 0.006$

$\sigma(5) = 0.99$

$\sigma(4) = 0.98$

$\sigma(3) = 0.95$

$\sigma(2) = 0.88$

$\sigma(1) = 0.73$

Properties of Sigmoid function

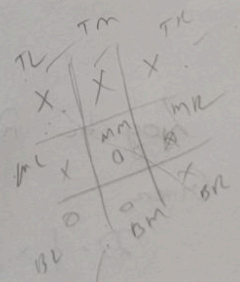
1) $0 < \sigma(z) < 1 \quad \forall z \quad -\infty \le z \le \infty$

2) $\sigma(z) = \frac{1}{2}$ for $z = 0$

3) $z \to +\infty$ $\quad \sigma(z)$ close to 1

4) $z \to -\infty$ $\quad \sigma(z)$ " " 0

5) if $\sigma(z) = \frac{1}{1+e^{-z}}$, $\quad 1 - \sigma(z) = \frac{e^{-z}}{1+e^{-z}}$



Derivative of Sigmoid fn

$$\sigma'(z) = \frac{1}{1+e^{-z}}$$

$$= \frac{0 + e^{-z}}{(1+e^{-z})^2}$$

$$= \frac{e^{-z}}{(1+e^{-z})^2}$$

$$\sigma'(z) = \frac{1}{(1+e^{-z})} \cdot \frac{e^{-z}}{(1+e^{-z})}$$

$$= \sigma(z)\,(1-\sigma(z))$$

derivative of Sigmoid fn can be written as sigmoid fn itself

In classification SSE gives the no of misclassification (as it is Binary)

If y is a Bernoulli R.V

$$f(y) = p^y (1-p)^{1-y} \qquad p = P(\text{Success})$$

$y=0 \quad f(y) = 1-p$ (failure)

$y=1 \quad f(y) = p$ (Success)

Logistic Regression

Given

$D = \{x_i, y_i\}_{i=1}^{n}$ Normal RV (Continuous)

$x_i \in \mathbb{R}^d \quad (x_i \sim N(\mu, \sigma))$ Gaussian

$y_i \in \{c_1, c_2\}$

$f : x_i \to y_i$

Standard Normal Distribution

$\rightarrow \frac{x-\mu}{\sigma}$

Objective

$\hat{f} : \mathbb{R}^d \to \{c_1, c_2\}$

Show that $\{0^n : n \text{ is prime}\}$ is not regular

Suppose L is regular. By Pumping lemma

$\exists m, \forall w$ with $|w| \geq m$, $\exists w = xyz$ such that $|xy| \leq m$, $|y| \geq 1$, $xy^k z \in L \; \forall \; k \geq 0$

for any m,

choose, $w_m = 0^p$ where p is prime & $p > m$.

Consider, $x = 0^i$

$y = 0^j$

$z = 0^{i-i-j}$, $xy^q z$.

$xy^q z$ where $q = |x| + |z| \in L$

$xy^q z \in L$

$|xy^q z|$ is prime

$|x| + |y^q| + |z|$ is prime

$|x| + |z| + q|y|$ is prime.

$q + q|y|$ is prime

$q(1+|y|)$ is prime,

Note : $1 + |y| \neq 1$

By pumping lemma,

$xy^0 z \in L$

$|xy^0 z|$ is prime

$|x| + 0|y| + |z|$ is prime

$|x| + |z|$ is prime

q is prime

$q \neq 1$

when $q \neq 1$ & $1 + |y| \neq 1$

then $q(1+|y|)$ is not prime,

$\therefore$ Contradicts.

show that $\{0^n : n \text{ is not prime}\}$ not regular.
Suppose L is regular, then L^c is also regular.
$L^c : \{0^n : n \text{ is prime}\}$ is regular
a contradiction.

$L = \{a^n b^n : n \geq 0\} \cup \{a^n b^m : n, m \geq 2\}$ is regular?

$L_1 \cup L_2$, if $L_1 \subseteq L_2$

$\Rightarrow L_2 \cup \{ab, \lambda\}$

- union of 2 regular language is regular.

$L = \{a^n b^n : n \neq 100\}$ is not regular.

$L_1 = \{a^n b^n : n \geq 0\}$

$L_2 = \{a^{100} b^{100}\}$ $(n = 100)$

$L_1 = L \cup L_2$

Since L_1 is regular (ASSUM)

L_2 is regular (finite)

contradiction. $\therefore L_1 = L \cup L_2$ is also regular.

ST $L = \{w : |w|_a = |w|_b\}$ is not regular.

$L_1 = \{a^n b^n : n \geq 0\}$

$L_2 = a^* b^*$

$L_1 = L \cap L_2$

for CA1

→ formal language, operation on languages, Find L^R, L^c, L^*,

→ give a regEx, defn of DFA, NFA, regEx. (X) → pumping lemma

→ diff bet NFA & DFA

→ how to construct DFA?

→ product automata (starting with a & ending with bb)

→ equivalence, minimization of DFA.

(X) conversion of NFA (with ε, w/o ε) to DFA

(X) DFA to regEx, regEx to DFA.

→ Ardens lemma (no, state elimination / any one method).

→ closure properties. (L_1 is reg, how to prove $L_1 \cup L_2$ is reg, L_1^c is reg, etc..)