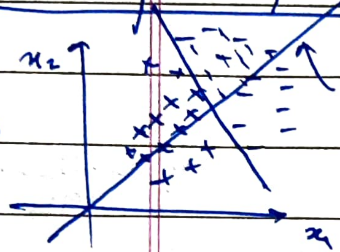
Week 8 - Feature Selection:

PAGE No.
DATE / /

5.1.) Motivation & creating new features:

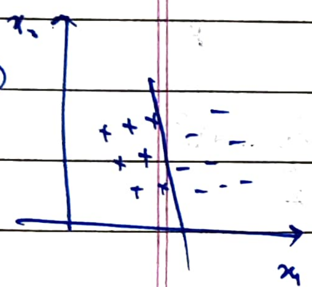
- Articulate why right features are important & wrong are harmful
- Be able to generate new ones from existing ones
- Be aware of common features for common data types
- Be able to apply feature selection methods.

Motivating Examples:

① 

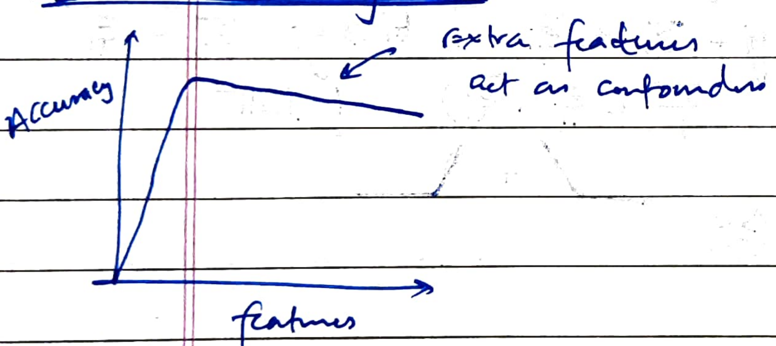
Any decision boundary of x_2 / x_1 parallel will misclassify.

But a projection line will seperate them.

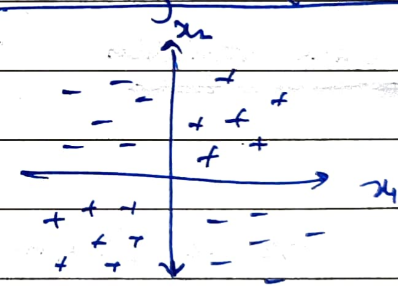
② 

Even though x_1 can classify on its own x_2 affects slant of the decision boundary & acts as a confounder.

So, we essentially have :-



Generating new features:-



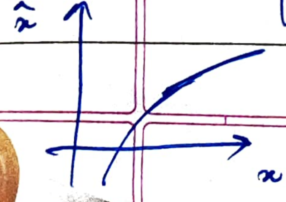
$x_3 = x_1 x_2$ plane elevates +ve samples

Transforming variables:

Power: $\hat{x} = x^p$ | log: $\hat{x} = \log(x+c)$ | exp: $\hat{x} = e^{ax+b}$

- Transforming big range ⟷ small range etc.

Objective is to transform to an easier distribution:



log for example

objective:

$\hat{x}$ should have an easier dist. in the analytical sense.

Lets say $\hat{x} = f(x)$ assuming they are injective

$\Rightarrow f^{-1}(\hat{x}) = x$

PAGE No.
DATE / /

If $p(x)$ and $q(\hat{x})$ are the distributions of the two,

$$q(\hat{x}) = p(f^{-1}(\hat{x})) \left| \frac{d}{dx} f^{-1}(\hat{x}) \right| \Rightarrow q(\hat{x}) = \frac{p(f^{-1}(\hat{x}))}{\left| \frac{d}{dx} f(\hat{x}) \right|}$$

Derivation :-

$$P(f(x) \leq \hat{x}) = P(x \leq f^{-1}(\hat{x})) = F(f^{-1}(\hat{x}))$$

Method of transformations:

$Y = g(x)$ g is differentiable | g is strictly increasing.

Now, $F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y))$
$= F_X(g^{-1}(y))$

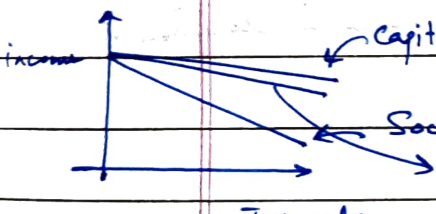
$$\Rightarrow \frac{d}{dy} F_Y(y) = f_Y(y) = \frac{d}{dy} F_X(g^{-1}(y)) = \frac{f_X(x_1)}{g'(x_1)} \quad [x_1 = g^{-1}(y)]$$

Dummy Variables:

- Convert discrete variables with N levels to N-1 binary variables.

Apple, Orange, Banana, Grapes $\Rightarrow W_1 x_{d_1} + W_2 x_{d_2} + W_3 x_{d_3} + b$ (4^{th})

- Interaction variables



Capit.

Socialist & Capitalist countries

On seperating.

Socialist

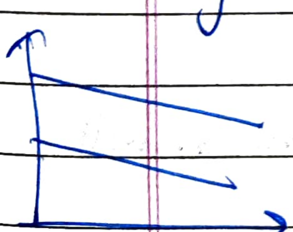
Average effect. (If we don't include type of economy)

Tax rate

$$y = W_1 x_1 + W_2 x_2 + b \quad (+ W_3 x_1 x_2)$$

social economy type

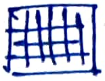
Models the original graph with different slopes.



If combination is important $\Rightarrow$ Interaction term needed.

x_1 x_2 $x_1 x_2$ are correlated and can have some effects.

5.2) Features for images:



Common Image Features:

- Pixel level statistics
  - Gray scale histogram
  - Color histograms
- Texture based stats.
  - Fourier descriptors
  - GLCM
    Grey level co occurrence Matrix
- Hybrid
- Shape based.

- Hu invariant moments $\ni \eta_{ij} = \sum \sum (x-\bar{x})^i (y-\bar{y})^j I(x,y)$

5.3) Features for audio & text.

- Meaningful features - power in diffrent frequency bands

Signal ⟶ fourier transform → Useful data.

Time windows also important. Window → Fourier Transform

MFCC is a default feature selection
1. Windows for time (overlapping)
2. DFT of window - Power Spectral Density
3. Filter bank - log energy
4. DCT (Discrete Cosine Transform)

Common Features to Extract from text:

"The cat was chasing the rat" or "I got lucky in the test today"
"I was not very lucky today"

- Histogram of words from a dictionary (feature vector) → $\begin{bmatrix} 0 \\ \vdots \\ \vdots \end{bmatrix}$ are not encoding.
  Problem: Too much importance to all words.
- TF-IDF: Term frequency - Inverse Document Frequency

$$tf(t,d) = \frac{f(t,d)}{\sum_{\tau \in d} f(\tau,d)}$$

Term frequency over a document

$$idf(t,D) = \log \frac{|D|}{1 + \{d \in D : t \in d\}}$$

D ← corpus

Entire feature = $tf \times idf$

(Also, pretrained deep neural network help for features as well)

5.4) Feature reduction 1 of 2 :

Features on pretrained deep neural networks :-

filter based methods / (Filter / Wrapper / Embedded)

x_{11} ... $x_{1,1000}$
⋮ ⋮
$x_{1000,1}$ ... $x_{1000,100}$

2^{100} subsets so too much

Filter based : For each feature decide keep vs discard → Train ML with kept features.

Wrapper : Generate subset, train & validate ML model on subset.

Embedded : - LASSO regularization, Subset selection & ML integrated.

Correlation - based elimination:

Blocks around diagonal - Feature reduction [Correlation based feature reduction]

Utility based subset selection:

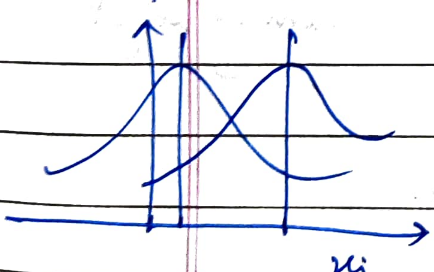
- Regression : Correlation | ◦ Classificⁿ : t-test | ◦ AIC & BIC

How the features related to target output.

1.) x_i correlation ⟹ Highly correlated with t may be good subset for prediction.

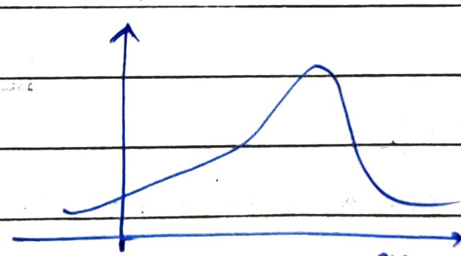
- Doesn't take in acc. interaction of x_i & x_j • Filter based method

2.) Classifⁿ : Two class classification.



$t=-1$

$t \in \{-1, 1\}$



$t=1$

Relative to own variances : Assumes they are distributed as Gaussian

t-test formula : $\left(\dfrac{\mu_1 - \mu_{-1}}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_{-1}^2}{n}}} \right)$ width of Gaussian important

AIC & BIC : Akai Info Criteria & Bayesian Info Criteria

5.5.) Forward Selection & Backward Elimination.

$x_1 \longleftrightarrow x_{100}$

Subset $S \leftarrow \{\phi\}$ initially $R - (x_1, \ldots, x_{100})$

for $i = 1 \to n$ (100) we measure the marginal utility.

for x_j in R

Measure marginal utility of x_j in ML model

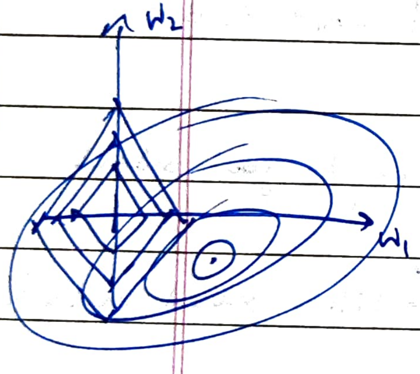
Include x_a with largest Marginal Utility in S & out of R.

Backward Eliminⁿ works in opp. dirⁿ :

Issues: If x_i & x_j correlated, we may not be able to find that

Backward elimination - correlated variables may already be in S

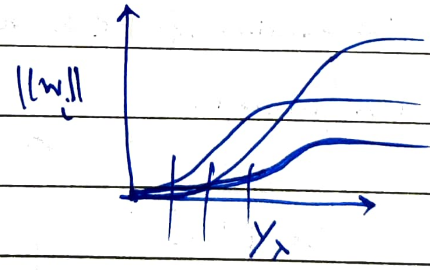
LASSO & elastic net:



Contours will graze to one of the corners.

$$L(w) = E(w) + \lambda \|w\|_1 + \lambda_2 \|w\|_2$$

Elastic Net



summing weights

Elastic Net :

keeps correlated variables kept in or out together.

Assume: $x_1 = x_2$

$w_1 x_1 + w_2 x_2 + w_3 x_3$ we want $w_1 \simeq w_2$

but $(w_1 + \alpha) x_1 + (w_2 - \alpha) x_2 + w_3 x_3$ is also the same

but L_1 norm is same. How to diff.? $\Rightarrow$ L2 Norm

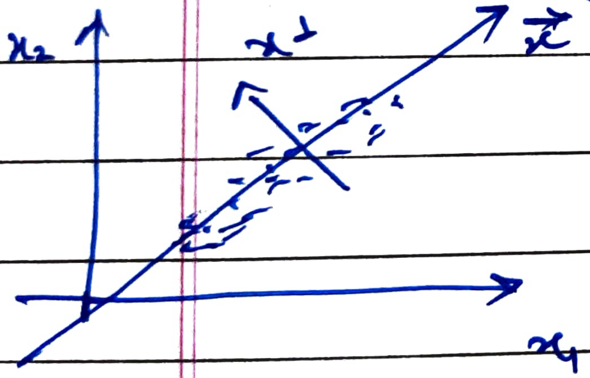
L2 penalty large for $(w_1 + \alpha)$ & $(w_2 - \alpha)$ case.

L1 SVM : (or L2 SVM is original SVM)

$$\frac{1}{2}\|w\|_2^2 + C\sum_i [1 - t_i y_i]_+$$

$$\lambda\|w\|_1 + C\sum_i [1 - t_i y_i]_+$$ for some values of λ you have certain w to be 0.

Principal Component Analysis :



($N \to d$)

Captures essence of x_1 & x_2.

Eigen vectors corresponding to higher Eigen values. Each captures contribution from all underlying contribution.