Linear Regression

This is a pretty old topic in the area of statistics and considered as a tool in supervised learning. (Work from Gauss)

MOTIVATION

Consider the following examples

(1) Predicting life time of an individual given body mass index

(2) Predicting crop yield given soil PH level, moisture and ambient temperature

(3) Predicting sales given advertising budget.

We are interested in predicting the quantitative response of a variable y given the variables $x_1, x_2, \ldots, x_n$

What we seek via models?

(1) Relationship between a variable to a quantitative response

(2) Assay the strength of the relationship & which variable contributes more..

(3) Accurately predict the future

There are other motivations as well!

Simple Linear Regression (1 - variable case)

$y \approx \alpha_0 + \alpha_1 x$

i.e., we are regressing y on x

α_0 : intercept

α_1 : slope

We need to estimate α_0 and α_1 from data

Estimating the coeffts

Let $\hat{\alpha}_0$ and $\hat{\alpha}_1$ be the estimates of the model parameters. To predict the future response $\hat{y}$ in response to variable x, we form

$$\hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1 x$$

Given: Data $\left\{ (x_i, y_i) \right\}_{i=1}^{n}$

Let us form the error for the data point (x_i, y_i) w.r.t. the predictor $\hat{y}_i$

$$\varepsilon_i = y_i - \hat{y}_i$$

$$= y_i - \underbrace{(\hat{\alpha}_0 + \hat{\alpha}_1 x_i)}$$

(deviation)

We formulate using the least square criterion

(Other criteria are possible!)

Consider the residual sum of squares (RSS)

$$RSS = \sum_{i=1}^{n} \varepsilon_i^2$$

$$RSS = \sum_{i=1}^{n} \Big(y_i - \underbrace{\hat{\alpha}_0 - \hat{\alpha}_1 x_i}_{\hat{y}_i}\Big)^2$$

<u>Goal</u>: $\min_{\hat{\alpha}_0, \hat{\alpha}_1} RSS$

We invoke basic calculus

Set $\dfrac{\partial RSS}{\partial \hat{\alpha}_0} = 0$; $\dfrac{\partial RSS}{\partial \hat{\alpha}_1} = 0$; Verify $\dfrac{\partial^2 RSS}{\partial \hat{\alpha}_i^2} > 0$

$i = 0, 1$

Taking derivatives

$$\frac{\partial RSS}{\partial \hat{\alpha}_0} = \underbrace{-2\sum_{i=1}^{n}\left(y_i - \hat{\alpha}_0 - \hat{\alpha}_1 x_i\right) = 0}$$

$$\Rightarrow \sum_{i=1}^{n} y_i = n\hat{\alpha}_0 + \hat{\alpha}_1 \sum_{i=1}^{n} x_i \quad \text{—— (A)}$$

Define $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$; $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ (sample mean)

(A) simplifies to $\boxed{\hat{\alpha}_0 = \bar{y} - \hat{\alpha}_1 \bar{x}}$ —— (B)

Now,

$$\frac{\partial RSS}{\partial \hat{\alpha}_1} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\alpha}_0 - \hat{\alpha}_1 x_i \right) x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = \hat{\alpha}_0 \sum_{i=1}^{n} x_i + \hat{\alpha}_1 \sum_{i=1}^{n} x_i^2 \qquad \text{—— (C)}$$

Using (B) in (C)

$$\sum_{i=1}^{n} x_i y_i = \left(\bar{y} - \hat{\alpha}_1 \bar{x} \right) \sum_{i=1}^{n} x_i + \hat{\alpha}_1 \sum_{i=1}^{n} x_i^2$$

Simplifying we get,

$$\hat{\alpha}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i} \qquad \text{— (D)}$$

Now let us simplify the numerator & the denominator terms to a compact form

OBSERVE

$$n\,\bar{y}\,\overbrace{\frac{1}{n}\sum_{i=1}^{n} x_i}^{\bar{x}} = n\,\bar{y}\,\bar{x}$$

|||ly $$n\,\bar{x}\,\overbrace{\frac{1}{n}\sum_{i=1}^{n} y_i}^{\bar{y}} = n\,\bar{y}\,\bar{x}$$

Also $$\sum_{i=1}^{n} \bar{x}\,\bar{y} = n\,\bar{y}\,\bar{x}$$

— (I)

Using (I) for simplifying numerator in (D)

$$\underbrace{\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i}_{\text{Numerator term in (D)}} = \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \bar{x}\,\bar{y}$$

$$= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

The numerator can be compactly written as

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

|||ly, let us consider the denominator

$$\sum_{i=1}^{n} x_i^2 - n\bar{x}\overbrace{\frac{1}{n}\sum_{i=1}^{n} x_i}^{\bar{x}} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$= \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \left(\because \sum_{i=1}^{n} x_i^2 - \overbrace{2n\bar{x}\frac{1}{n}\sum_{i=1}^{n} x_i}^{2n\bar{x}^2} + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right)$$

Compact form

Writing it compactly

$$\hat{\alpha}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Typically, we may not know the true relationship of x with y

('Assume' ε is independent of x) with mean zero

$$y = f(x) + \varepsilon$$

random term (noise)

To assess the accuracy of the fit, you need to evaluate the model error

Also, errors ε_i may be correlated. These have to be taken into account appropriately.

Consider $E(y-\hat{y})^2$ Some model for $\hat{f}$

a: $f(x)-\hat{f}(x)$

b: ε

$$= E\left(\underbrace{f(x)+\varepsilon}_{\text{}} - \hat{f}(x)\right)^2$$

$$= E\left[\left(\underbrace{f(x)-\hat{f}(x)}\right)^2\right] + E(\varepsilon^2) + \underbrace{\cancel{2E\left(f(x)-\hat{f}(x)\right)E(\varepsilon)}}_{\text{zero}} \to 0$$

$$= \underbrace{E\left[\left(f(x)-\hat{f}(x)\right)^2\right]}_{\text{Can optimize based on choice of } \hat{f}} + \underbrace{\mathrm{var}(\varepsilon)}_{\text{Cannot reduce this error}}$$

In the noiseless case, $y = f(x)$ & $\hat{f}(x) = \alpha_0 + \alpha_1 x$

Maximum likelihood estimation for the linear regression model

Suppose we have data points (x_i, y_i); $i = 1, \dots, n$

Consider the model $y_i = f(x_i) + \varepsilon_i$

$\varepsilon_i \sim N(0, \sigma^2)$

and (x_i, y_i) are iids (independent and identically distributed)

Our linear regression model implies

$$\hat{y}_i = \alpha_0 + \alpha_1 x_i$$

Let $\underline{\alpha} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix}_{2\times 1}$ $\underline{x}_i = \begin{bmatrix} 1 \\ x_i \end{bmatrix}_{2\times 1}$; $\underbrace{\hat{y}_i = \underline{\alpha}^T \underline{x}_i}_{\text{compact form}}$

$$L(\underline{\alpha}) = \prod_{i=1}^{n} P(y_i \mid \underline{x}_i)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(\overbrace{y_i - \hat{y}_i}^{\varepsilon_i}\right)^2\right)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \underbrace{\underline{\alpha}^T \underline{x}_i}_{\hat{y}_i}\right)^2\right)$$

We take the logarithm of the likelihood fn since $\log(\cdot)$ is monotonic

Constant term

$$\ell(\underline{\alpha}) = -\frac{1}{2}\sum_{i=1}^{n}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \underline{\alpha}^T \underline{x}_i\right)^2$$

log likelihood

We set $\dfrac{\partial \ell(\underline{\alpha})}{\partial \underline{\alpha}^T} = 0$

To max. the log likelihood,

we need to minimize the term J

$$J = \sum_{i=1}^{n} \left(y_i - \underline{\alpha}^T \underline{x}_i\right)^2$$

Interpret $e_i = y_i - \underline{\alpha}^T \underline{x}_i$ scalar

$$\underline{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}_{n\times 1}; \quad \underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}; \quad X = \begin{bmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_n^T \end{bmatrix}_{n\times 2}$$

$$J = \underline{e}^T \underline{e} = (\underset{n\times 1}{\underline{y}} - \underset{n\times 2}{X} \underset{2\times 1}{\underline{\alpha}})^T (\underline{y} - X\underline{\alpha})$$

$$J = (\underline{y} - X\underline{\alpha})^T (\underline{y} - X\underline{\alpha})$$

$$J = (\underbrace{\underline{y}^T\underline{y}}_{\text{ignore}} - \underbrace{\underline{y}^T X\underline{\alpha} - \underline{\alpha}^T X^T\underline{y}}_{-2(X\underline{\alpha})^T\underline{y}} + \underline{\alpha}^T X^T X\underline{\alpha})$$

$$\frac{\partial J}{\partial \underline{\alpha}^T} = 0 \Rightarrow \boxed{-2\underset{2\times n}{X^T}\underset{n\times 1}{\underline{y}} + 2\underset{2\times n}{X^T}\underset{n\times 2}{X}\underset{2\times 1}{\underline{\alpha}} = 0}$$

$$\Rightarrow X^T X\underline{\alpha} = X^T\underline{y}$$

$$\underline{\alpha} = \underbrace{(X^T X)^{-1}}_{\text{if it exists!}} X^T\underline{y}$$

Multi variable linear regression (> 1 variable)

Suppose we have more than 1 variable, say p

'p' predictors (variables)

$$y \approx \alpha_0 + \alpha_1 x + \alpha_2 x_2 + \dots + \alpha_p x_p$$

Set up RSS as

$$RSS = \sum_{i=1}^{n} (\underbrace{y_i}_{\text{true observation}} - \underbrace{\hat{y}_i}_{\text{model}})^2$$

(Least Squares Criterion)

$$RSS = \sum_{i=1}^{n} \Big(y_i - \underbrace{\hat{\alpha}_0 - \hat{\alpha}_1 x_{i1} - \cdots - \hat{\alpha}_p x_{ip}}_{\hat{y}_i} \Big)^2$$

Choose $\hat{\alpha}_0^* \ldots \hat{\alpha}_p^* = \min\limits_{\hat{\alpha}_0 \ldots \hat{\alpha}_p} RSS$

NOTE: Solving the RSS optimization problem exactly can be tricky due to simultaneous eqns involved

One approach to tackle this is by Gradient descent technique

$$RSS = J(\underline{\hat{\alpha}}) = \sum_{i=1}^{n} \left(y_i - \underline{\hat{\alpha}}^T \underline{x}_i\right)^2$$

where $\underline{x}_i = \begin{bmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}$ $\quad \underline{\hat{\alpha}} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix}$

Update rule

$$\underline{\hat{\alpha}}_t = \underline{\hat{\alpha}}_{t-1} - \eta \left. \nabla J(\underline{\hat{\alpha}}) \right|_{\underline{\hat{\alpha}}_{t-1}}$$

update step

learning rate

Features need not be on the same scale

For grad. descent to work well

(1) We need to do feature scaling

(2) Do mean normalization (Features having zero mean)

Example for feature scaling :

x_1 : 0 – 120 years age $\Rightarrow \tilde{x}_1 = \frac{x_1}{120}$

x_2 : 0 – 7 children $\Rightarrow \tilde{x}_2 = \frac{x_2}{7}$

For mean normalization, replace each x_i with $x_i - \mu_i$. (This does not apply for $x_{0i} = 1$ case)

Important Qns

1) Do all the variables help in predicting y?
2) How well does the model fit?
3) Given predictor values, what response value do we predict? Is it a good prediction?

Deciding on the dominant variables

Practical Considerations

Real life data will require a subset of predictors to fit the quant. response.

How do we choose the best model

Example: Say $p = 2$ i.e., 2 predictors

a) Model with x_1 alone

b) —"— x_2 alone

c) —"— x_1 and x_2

d) No variable

$\boxed{2^p \text{ choices}}$

Practical Heuristics

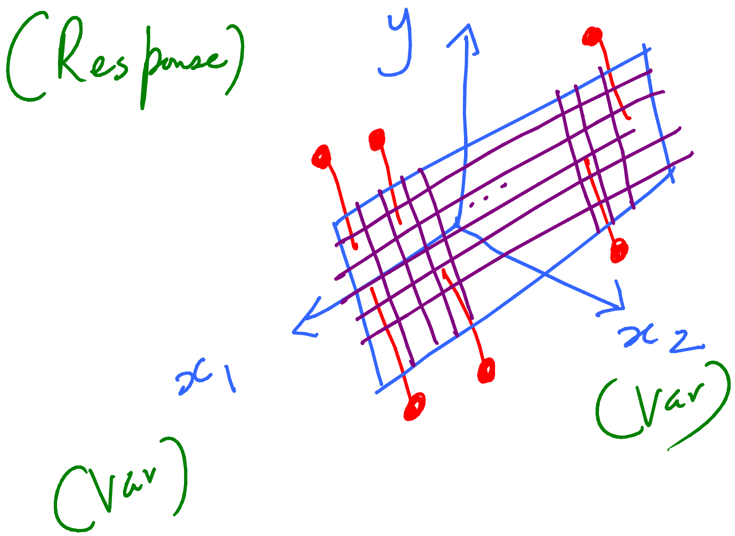
1) Forward Selection ; Start with a null model.
Fit p simple linear regressions (i.e., 1- variable case) & add to the null model the variable that gives the least RSS. To this, proceed with the variable with lowest RSS over a new 2- variable model etc.
Sequentially

2) Backward Selection:

We can proceed with all the Variables to start with and remove the variable which is least statistically significant i.e., peel off the variables sequentially

NOTE: Mixed approaches are also possible!

Visualization over a 2-variable case



Overall choices in the x_1, x_2 plane, y is observed

We need the optimal parameters for the eqn of the plane to fit the observations

Other variants

I. Using indication variables

Suppose $x_i = \begin{cases} 1 & i^{th} \text{ person has } IQ > 160 \\ 0 & \text{else} \end{cases}$

We can use such variables as predictors in the regression eqn

$$y_i = \underbrace{\alpha_0 + \alpha_1 x_i + \varepsilon_i}_{} = \begin{cases} \alpha_0 + \alpha_1 + \varepsilon_i & i \in IQ > 160 \\ \alpha_0 + \varepsilon_i & \text{else} \end{cases}$$

Indication Variable

Extensions to linear models

Std. linear regression makes 2 important assumptions between predictors and responses.

(a) Additive assumption:
Effect of change in predicting x_i on y is independent of the rest $x_{j \setminus i}$

(b) Linearity
Change in response to 1 unit change in x_i is constant, regardless of x_i

Can we relax the additive assumption?

Idea: Include 'interaction' terms

Suppose $y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \varepsilon$

(Here 1 unit change in x_1 say, $\alpha_1 \uparrow$)

If $y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \underbrace{\alpha_3 x_1 x_2} + \varepsilon$

$= \alpha_0 + \underbrace{(\alpha_1 + \alpha_3 x_2)} x_1 + \alpha_2 x_2 + \varepsilon$

$= \alpha_0 + \alpha_1' x_1 + \alpha_2 x_2 + \varepsilon$

Effect of x_1 on y is no longer constant!

Adjusting x_2 influences x_1 on y.

Example : Imagine an assembly line in a manufacturing unit

Let x_1 : # production lines $\quad y$: # units manufactured

x_2 : # workers

If # workers = 0, increasing x_1 will not yield y

i.e., $x_2 = 0$

$$\#\text{Units} \approx \alpha_0 + \alpha_1 \#\text{prod.-lines} + \alpha_2 \#\text{workers} + \underbrace{\alpha_3 (\#\text{prod.-lines} \times \#\text{workers})}_{\text{Interactions}}$$

Other Issues

Suppose $y = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \varepsilon$ — ①

Constant accelaration

Eg: displacement $S = ut + \frac{1}{2} a t^2$

non-linear fn of time 't'

Constant initial velocity

variable

Define $x_1 : x$

$x_2 : x^2$

$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \varepsilon$ — I

Note: Eqn I is still a multiple variable linear regression model

Issues to Consider

1) Non linear relationships between variables & response
2) Correlation of errors
3) Outliers
4) Collinearity of 2 or more variables

⋮

Logistic Regression

MOTIVATION

There are scenarios requiring qualitative responses. In such cases, linear regression may not be the right choice

Example: Suppose we are trying to predict the condition of a crop with diagnosis as

qualitative responses

(a) excessive manure (b) pest issues

(c) low moisture etc. based on a set of predictors x_1 x_2 ... x_p

Let us form a quantitative response using the foll. encoding

$$y = \begin{cases} 1 & \text{excessive manure} \\ 2 & \text{pest issues} \\ 3 & \text{low moisture} \end{cases}$$

One can do a Least squares (LS) fit to a linear regression model based on $x_1, x_2, \dots x_p$

However, we can have a different encoding rule

$$y = \begin{cases} 1 & \text{pest issues} \\ 2 & \text{low moisture} \\ 3 & \text{excessive manure} \end{cases}$$

Note that a different encoding rule can give a totally different relationship to the conditions

$\Longrightarrow$ We have fundamentally different models leading to different set of predictors

If the qualitative response variable has a natural ordering

E.g., mild spicy, medium spicy, hot spicy etc.

A coding scheme of 1, 2, 3 in that order is reasonable

Note that for a binary response i.e., 0/1 encoding there is no problem since

$$y = \begin{cases} 1 & \text{Case } A \\ 0 & \text{Case } B \end{cases} \implies \begin{array}{l} \hat{y} > 0.5 \implies \text{Case } A \\ \hat{y} \leq 0.5 \implies \text{Case } B \end{array}$$

This motivates us to develop classification methods suited for qualitative responses

---

LOGISTIC REGRESSION is one such method.

Logistic Regression

Applications (Examples)

1) Predicting failure of a product given indicators / predictors.
2) Predict if a home owner defaults on a loan given a bank balance history etc.

⋮

In the simple linear regression case,

$$p(x) = \alpha_0 + \alpha_1 x$$

Depending on the value of x, one can have $p(x) < 0$ or $p(x) > 1$.

However, if we want to interpret $p(x)$ in terms of a probability, we need to limit $p(x)$ between 0 and 1

(Cond. Prob. of y given x) i.e., $P : \mathbb{R} \longrightarrow (0,1)$

observable

One such fn is the "logistic function"

$$p(x) = \frac{e^{\alpha_0 + \alpha_1 x}}{1 + e^{\alpha_0 + \alpha_1 x}}$$

logistic function

$$\frac{p(x)}{1 - p(x)} = e^{\alpha_0 + \alpha_1 x};$$

Interpret this as odds that can take values $\in (0, \infty)$

Taking logs.

$$\ln\left(\frac{p(x)}{1 - p(x)}\right) = \alpha_0 + \alpha_1 x$$

(Linear function)

1 unit $\uparrow$ $x \Rightarrow \alpha_1 \uparrow$ in logit

log. odds or logit

There are a few points to note

1) There is no linear relationship between $p(x)$ and x

2) Rate of change in $p(x)$ per unit change in x depends on the current value of x.

With the set up of the model, our next step is to estimate the regression coeffts.

Estimating the regression coeffts

We shall shift gears on our metric than the RSS adopted in linear regression and use maximum-likelihood approach

Formulate the likelihood function

$$L(\alpha_0, \alpha_1) = \prod_{i : y_i = 1} p(x_i) \prod_{j : y_j = 0} \left(1 - p(x_j)\right)$$

$$= \prod_{i=1}^{n} p(x_i)^{y_i} \left(1 - p(x_i)\right)^{1 - y_i} \quad \text{Ⓐ} \quad (y_i \in \{0, 1\})$$

GOAL: Choose $\hat{\alpha}_0^*, \hat{\alpha}_1^* = \max\limits_{\alpha_0, \alpha_1} L(\alpha_0, \alpha_1)$

Since $\log(\cdot)$ is a monotonic fn, we take $\log(\cdot)$ of the likelihood function

$$\ell(\alpha_0, \alpha_1) \triangleq \log\left[L(\alpha_0, \alpha_1)\right] \text{ ——— } (B)$$

Simplifying (A) using (B)

$$\ell(\alpha_0, \alpha_1) = \sum_{i=1}^{n} \left[y_i \log p(x_i) + (1 - y_i) \log\left(1 - p(x_i)\right) \right]$$

$$= \underbrace{\sum_{i=1}^{n} \log\left(1 - p(x_i)\right)}_{\text{1st term}} + \underbrace{\sum_{i=1}^{n} y_i \log\left(\frac{p(x_i)}{1 - p(x_i)}\right)}_{\text{2nd term}}$$

$$= -\sum_{i=1}^{n} \log\left(1 + e^{\alpha_0 + \alpha_1 x_i}\right) + \sum_{i=1}^{n} y_i \underbrace{(\alpha_0 + \alpha_1 x_i)}_{\text{linear term in the logit.}}$$

The usual way is to take $\dfrac{\partial \ell(\alpha_0, \alpha_1)}{\partial \alpha_0} = 0$

and $\dfrac{\partial \ell(\alpha_0, \alpha_1)}{\partial \alpha_1} = 0$. Verify that $\dfrac{\partial^2 \ell(\cdot)}{\partial \alpha_i^2} < 0$ $(i = 0, 1)$

for max.

Taking the partial derivatives & setting = 0

$$\frac{\partial l(\cdot)}{\partial \alpha_0} = \underbrace{-\sum_{i=1}^{n} \frac{e^{\alpha_0+\alpha_1 x_i}}{1+e^{\alpha_0+\alpha_1 x_i}}}\ \underbrace{+\sum_{i=1}^{n} y_i} = 0 \quad \text{C}$$

$$\frac{\partial l(\cdot)}{\partial \alpha_1} = \underbrace{-\sum_{i=1}^{n} \frac{e^{\alpha_0+\alpha_1 x_i}}{1+e^{\alpha_0+\alpha_1 x_i}} \cdot x_i}\ \underbrace{+\sum_{i=1}^{n} y_i x_i} = 0 \quad \text{D}$$

Eqns. C and D cannot be solved in closed form (transcendental equations) Need numerical evaluations est. $\to \hat{\alpha}_0^* \leftarrow$ opt.

Once we get the opt. estimates $\hat{\alpha}_0^*$ $\hat{\alpha}_1^*$, we can predict the response

$$\hat{p}(x) = \frac{e^{\hat{\alpha}_0^* + \hat{\alpha}_1^* x}}{1 + e^{\hat{\alpha}_0^* + \hat{\alpha}_1^* x}}$$

i.e., Plug in x, compute $\hat{p}(x)$ and predict the qualitative decision for e.g., defaulting = Yes/No over a home loan given the bank balance.

Binary response to multiple predictors

Applns :

1) Would I go for pure science or engg. for my under grad given (a) my grades (b) likes/dislikes?

2) Which of the 2 parties will an individual vote given (a) demographic characteristics (b) likes/dislikes?

etc.

⋮

Plenty of examples

From our ideas earlier,

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x_p$$

$x_1, x_2, \ldots, x_p$ are predictors

$$p(x) = \frac{e^{\alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x_p}}{1 + e^{\alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x_p}}$$

$$L(\alpha_0 \dots \alpha_p) = \prod_{i=1}^{n} \underbrace{p(x_1^{(i)} \dots x_p^{(i)})^{y_i}}_{} \left(1 - p(x_1^{(i)} \dots x_p^{(i)})\right)^{1-y_i}$$

Choose $\underbrace{\hat{\alpha}_0^* \dots \alpha_p^*}_{} = \max\limits_{\alpha_0 \dots \alpha_p} L(\alpha_0, \dots, \alpha_p)$

Generalization to the K-class problem

Consider the linear predictor with p predictors leading to outcome 'k' $(k = 1, \ldots, K)$ i.e., observation 'i'

reg coefft · pred. variable

Let $\phi(k, i) = \alpha_{0,k} + \alpha_{1,k}\, x_{1,i} + \cdots + \alpha_{p,k}\, x_{p,i}$

observation

Each coefft $\alpha_{j,k}$ is the regression coefft.

In $\alpha_{j,k}$; $j = 0, \ldots p$ (reg. coeffts)

Writing it compactly in vector form

$$\phi(k,i) = \underline{\alpha}_k^T \, \underline{x}_i$$ Compact form!

where $\underline{\alpha}_k = \begin{bmatrix} \alpha_{0,k} \\ \vdots \\ \alpha_{p,k} \end{bmatrix}$ $\underline{x}_i = \begin{bmatrix} 1 \\ x_{1,i} \\ \vdots \\ x_{p,i} \end{bmatrix}$

Interpreting the problem as independent binary regressions

We set one of the outcomes as a "pivot" and rest $K-1$ are regressed w.r.t the pivot

$$\ln\left(\frac{P(y_i=1)}{P(y_i=K)}\right) = \underline{\alpha}_1^T \underline{x}_i$$

pivot

$$\vdots$$

$$\ln\left(\frac{P(y_i=K-1)}{P(y_i=K)}\right) = \underline{\alpha}_{K-1}^T \underline{x}_i$$

Notational Use

$P(y_i=1)$ MUST be interpreted as $P(y_i=1 \mid \underline{x}_i)$

Conditioned to $\underline{x}_i$ i.e., predictors

For ease of notation, we use $P(y_i=1)$

Now,

$$P(y_i = 1) = P(y_i = K)\, e^{\underline{\alpha}_1^T \underline{x}_i}$$

$$\vdots$$

$$P(y_i = K-1) = P(y_i = K)\, e^{\underline{\alpha}_{K-1}^T \underline{x}_i}$$

Since $P(y_i = K) = 1 - \sum_{k=1}^{K-1} P(y_i = k)$

$$\left(\sum_{k=1}^{K} P(y_i = k) = 1 \right)$$

Prob. sum.

$$\boxed{P(y_i = K) = \frac{1}{1 + \sum_{j=1}^{K-1} e^{\underline{\alpha}_j^T \underline{x}_i}}}$$

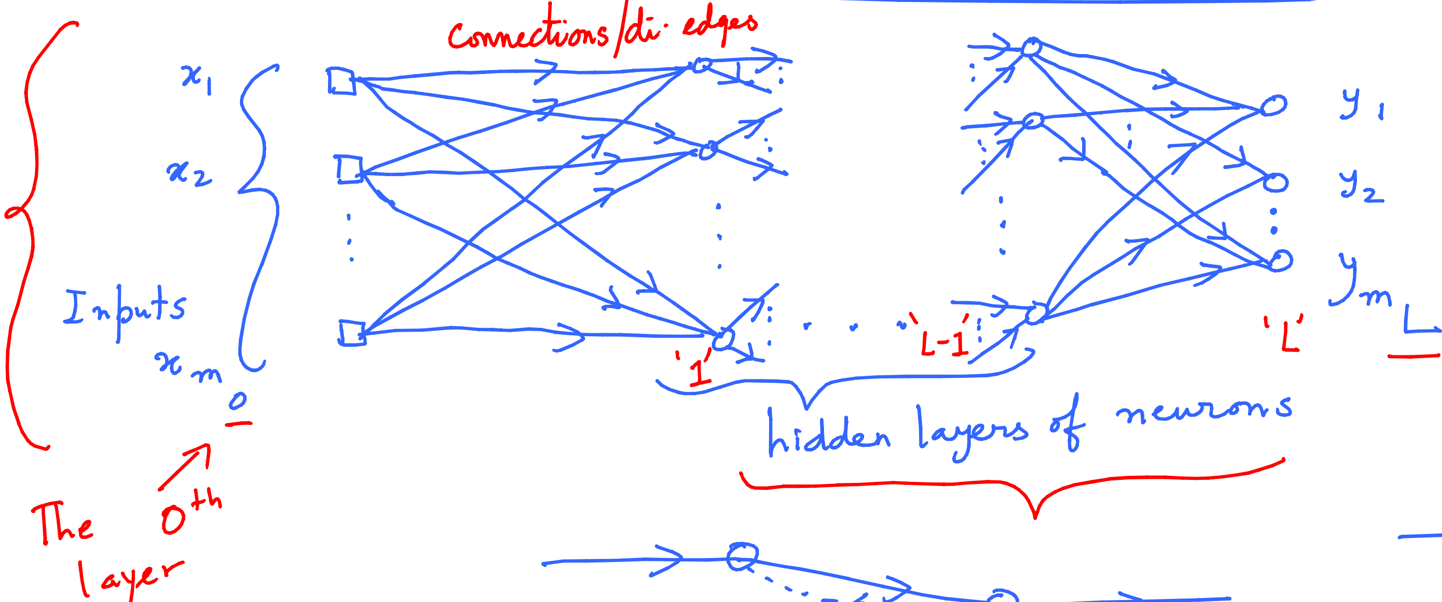
$$\Rightarrow \quad P(y_i = j) = \frac{e^{\underline{\alpha}_j^T \underline{x}_i}}{1 + \sum_{k=1}^{K-1} e^{\underline{\alpha}_k^T \underline{x}_i}}$$

Prob. out come $= j$ for observation 'i'

One can proceed towards estimating $\{\underline{\alpha}_k\}_{k=1}^{K}$ using maximum aposteriori probability criterion.

Ref: Refer to any stat. modeling book for a more advanced reading on the material.

Multilayer Perceptron



← The last layer

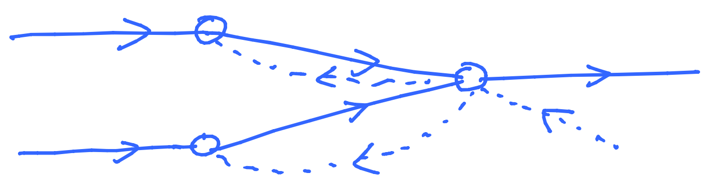
Error

$$\underline{e} = \underline{y} - \underline{d}$$

Legend

→ functional signals

- - <- - error signals



Neuron : non-linear processing element

fully connected network
from every node/neuron
to every other node/neuron

Let $y_j(n)$ denote the function signal at the o/p of neuron j in the o/p layer to a stimulus $\underline{x}(n)$ @ the input

desired attribute/coordinate of $\underline{d}$

$$e_j(n) = d_j(n) - y_j(n)$$

where $d_j(n)$ is the j^{th} element of $\underline{d}(n)$

The instantaneous error energy

$$\underbrace{\mathcal{E}_j(n)}_{\text{discrete time instants 'n'}} = \underbrace{\frac{1}{2}}_{\text{normalization}} \underbrace{e_j^2(n)}_{\text{Sq. error}}$$

Over all neurons in the o/p layer

$$\mathcal{E}(n) = \frac{1}{2} \sum_{j \in C} \underbrace{e_j^2(n)}_{\text{instantaneous energy for neuron 'j'}}$$

C is the set of neurons in the o/p layer

Now, with N training samples, we can average the error energy

$$\mathcal{E}_{av}(N) = \underbrace{\frac{1}{2N} \sum_{n=1}^{N} \sum_{j \in C} e_j^2(n)}_{\frac{1}{N} \sum_{n=1}^{N} \mathcal{E}(n)}$$

We have 2 modes

1) Batch Learning : Adjustments to the synaptic weights are performed after all N datapoints in the training set are presented to the n/w. Synaptic wts. are adapted on an epoch-by-epoch basis.

2) Online Learning :

Adjust weights for every tuple $(\underline{x}(i), \underline{d}(i))$ presented to the n/w @ time 'i'.

feature vector

desired response vector

PROS and CONS

Batch Learning

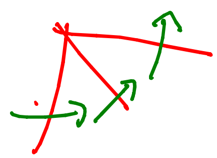
PROS

1) Accurate estimation of the gradient vector towards convergence
2) Parallelization of the learning process

CONS

Demanding on the storage requirements

$$\frac{1}{N}\sum_{n=1}^{N} \mathcal{E}(n)$$



Online Learning

PROS

1) Track small changes in the training data.
2) Make use of redundancy in the data set.
3) Easy to implement

Single most Good reason

CONS

Parallelization is not possible

Need to do ensemble averaging over large initial conditions.