Recap.
1. D-Trees
2. N-Bayes.

Linear Models

- Naive Bayes "under certain conditions" is a linear model.
  - (-) So, you learn a set of weights.
- Perceptrons, "first ANN", a linear classifier model.

Linear Models :-

Basic model :-

$$w_1 \bar{x}_1 + \cdots + w_k \bar{x}_k - w_0 \gtrless 0. \quad \text{(Perceptron)}$$

Question :- How to learn the weights?

already seen two ways of doing it.

- Can be posed as an optimization problem.

$$w_{OPT} = \arg \min_{\bar{w}, w_0} J(\bar{w}, T, w_0)$$

$$J(w_0, \ldots, w_k, T)$$

① detour to # mistakes

② P3

Choosing a different 'F' gives a different w^*_{OPT}

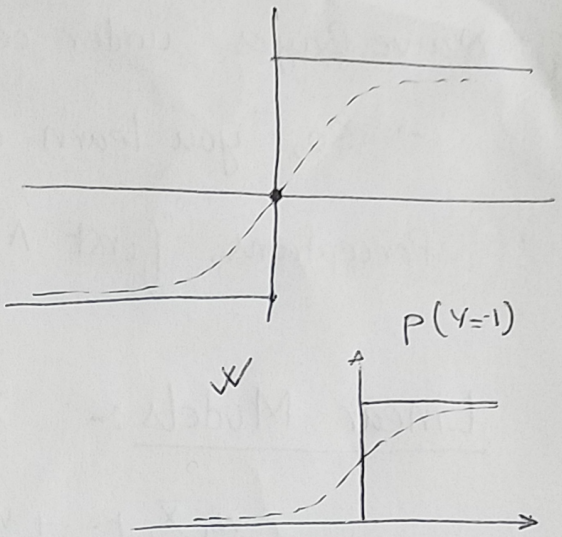
NB :- $P(x|Y), P(Y) \longmapsto P(Y|x)$.

Logistic Regression :- Come up with a model of $P(Y|x)$.

$$\text{Prob}(Y=0|\bar{x}) = \frac{1}{1+\exp(\bar{w}\cdot\bar{x}+w_0)}.$$

and,

$$P(Y=0|\bar{x}) = \frac{\exp(w_0 + \bar{w}\cdot\bar{x})}{1+\exp(w_0+\bar{w}\cdot\bar{x})}.$$



Classification rule:-

$$P(Y=1|\bar{x}) \gtrless P(Y=0|\bar{x})$$

$$\Leftrightarrow \exp(\bar{w}^*\cdot\bar{x}^*) \gtrless 1$$

$$\Leftrightarrow \boxed{\bar{w}\cdot\bar{x}+w_0 \gtrless 0}$$

how to train our logistic regression function?

$$\bar{w} \leftarrow \arg\max_{\bar{w}} \hat{\prod_{i=1}} P(f(\bar{x}(i)) \mid \bar{x}(i), \bar{w}^*)$$

For simplicity, we will take 'f' to be $\{+1, 0\}$ instead of $\{+1,-1\}$. $\boxed{(1+Y)(1-Y)=0}$

$Y \in \{+1,0\}$, then, $\boxed{Y\cdot(1-Y)=0}$

$$P(f(\bar{x}) \mid \bar{x}, \bar{w}) = \left(P(f(\bar{x})=1 \mid \bar{x}, \bar{w})\right)^{f(\bar{x})} \cdot \left(P(f(\bar{x})=0 \mid \bar{x}, \bar{w})\right)^{(1-f(\bar{x}))}$$

⇒ Log probability of the training examples given $\bar{w}$:-

$$l(\bar{w}) = \sum_{i=1}^{n} f(\bar{x}(i)) \cdot \log \text{Prob}\left(f(\bar{x}(i))=1 \mid \bar{w}, \bar{x}(i)\right) + (1 - f(\bar{x}(i))) \cdot \log \text{Prob}\left(f(\bar{x}(i))=0 \mid \bar{x}, \bar{w}\right) \quad (5.3)$$

$$= \sum_{i=1}^{n} f(\bar{x}(i)) \log \frac{P(f(\bar{x}(i))=1 \mid \bar{w}, \bar{x}(i))}{P(f(\bar{x}(i))=0 \mid \bar{w}, \bar{x}(i))} + \log P\left(f(\bar{x}(i))=0 \mid \bar{x}, \bar{w}\right)$$

$$= \sum_{i=1}^{n} f(\bar{x}(i)) \cdot \left(w_0 + \sum_{j=1}^{K} w_j \bar{x}_j(i)\right) - \log\left(1 + \exp\left(w_0 + \sum_{j=1}^{K} w_j x_j(i)\right)\right)$$

NO CLOSED FORM APPROACH/SOL

BUT:- unique solution ⇐ Concave function.

$\bar{w} \longrightarrow \log(1+\exp(\bar{x}\cdot\bar{w})) \longrightarrow$ convex function of $\bar{w}$.

Gradient ascent methods :-

$$\frac{\partial l(\bar{w})}{\partial w_i} = \sum_{i=1}^{n} f(\bar{x}(i)) \cdot \bar{x}_j(i) - \frac{\exp(\bar{w}_0^* \bar{x}^*)}{P(f(\bar{x}(i))=1 \mid \bar{x}, \bar{w})} \cdot \bar{x}_j(i)$$

$$= \sum_{i=1}^{n} \bar{x}_j(i) \cdot \left(f(\bar{x}(i)) - P(f(\bar{x}(i))=1 \mid \bar{x}, \bar{w})\right)$$

$$\therefore \quad \bar{w}^{(l)} = \bar{w}^{(l-1)} + \eta \cdot \nabla_{\bar{w}} l(\bar{w})$$

$$\bar{w}_j^{(l)} = \bar{w}_j^{(l-1)} + \eta \cdot \frac{\partial l(\bar{w})}{\partial w_i}$$

$\eta \longrightarrow$ large (overfitting)

$\eta \longrightarrow$ small (underfitting).

If prob close to '1' not much effect.

- ⊖ A natural measure :-

Find the $\bar{w}$ that minimizes the # of errors.

$$\bar{w}^*_{opt} = \arg\min_{\bar{w}^*} \left\{ \#\text{examples s.t. } f(\bar{x}(i)) \neq \text{sign}(\bar{w}^* \cdot \bar{x}(i)) \right\}$$

$$= \text{"} \quad \text{"} \quad \sum_{i=1}^{n} \mathbb{I}\left\{ f(\bar{x}(i)) \neq \text{sign}(\bar{w}^* \cdot \bar{x}^*(i)) \right\}.$$

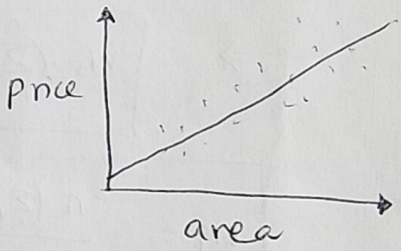
hard to compute.

OPTIMIZATION.

minimize (OR maximize) some $J(w_0, w_1, \ldots, w_k, T)$

↓ parameters to tune.

Simple example :- Regression (Linear).



- Output $f(\bar{x}(i))$ is a real value.

Suppose only one feature, w_1, $\bar{x}_1 = x$.

Area = x $\xrightarrow{\text{guess}}$ $w_0 + w_1 x$.

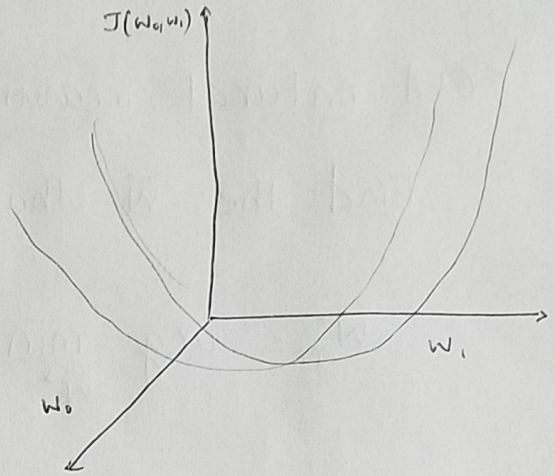
$(x_1, f_1), (x_2, f_2), \ldots, (x_n, f_n)$.

Fit a line :- $y = h(x) = w_0 + w_1 x$.

Least squared error :- $\sum_{i=1}^{n} (f_i - (w_0 + w_1 x))^2$

Find w_0, w_1 to minimize the error.

- Convex functions :- unique minima.



$$\frac{\partial}{\partial w_0} J(w_0, w_1, T) = 0$$

$$\Rightarrow 0 = \sum_{i=1}^{n} 2\left(f_i - (w_0 + w_1 x_i)\right) \cdot (-1)$$

$$\Rightarrow \sum_{i=1}^{n} f_i = \sum_{i=1}^{n} (w_0 + w_1 x_i)$$

$$\Rightarrow \boxed{n \cdot w_0 + w_1 \cdot (\Sigma x_i) = (\Sigma f_i)}$$

$$\frac{\partial}{\partial w_1} J(w_0, w_1, T) = 0 \Rightarrow 0 = \sum_{i=1}^{n} 2\left(f_i - (w_0 + w_1 x_i)\right) \cdot (+x_i)$$

$$\Rightarrow \boxed{w_0 \cdot \left(\sum_{i=1}^{n} x_i\right) + w_1 \cdot (\Sigma x_i^2) = (\Sigma f_i x_i)}$$

$$\boxed{w_1 = \frac{n \cdot (\Sigma f_i x_i) - (\Sigma f_i)(\Sigma x_i)}{\left(n \cdot (\Sigma x_i^2) - (\Sigma x_i)^2\right)}}$$

$$\boxed{w_0 = \frac{\Sigma f_i - w_1 (\Sigma x_i)}{n}}$$

Multivariate regression :- More features than just '1'

eg, area, # bedrooms, walk-score.

$$f = w_0 + w_1 \bar{X}_1 + w_2 \bar{X}_2 + \cdots + w_k \bar{X}_k = \boxed{\bar{w} \cdot \bar{x}^*} = \bar{w}^T \bar{x}^*$$

$$\boxed{J(\bar{w}^*, T) = \sum_{i=1}^{n} \left(f_i - \bar{w}^{x^+} \cdot \bar{X}(i)\right)^2} \longrightarrow \bar{w}^* = (X^T X)^{-1} X^T$$

$$\min_{\bar{w}^*} \underbrace{\|F - \bar{\bar{X}} \cdot \bar{w}^*\|_2^2}$$

$F \longrightarrow n \times 1$

$\bar{\bar{X}} \longrightarrow n \times (k+1)$

$\bar{w}^* \longrightarrow (k+1) \times 1$

$$= \sum_{i=1}^{n}$$

~~$F^T - \bar{w}^{*T}(\bar{\bar{X}}^T \bar{\bar{X}}) \cdot w$~~

$$\underbrace{(F^T - \bar{w}^{*T} \cdot \bar{\bar{X}}^T) \cdot (F - \bar{\bar{X}} w)} = F^T F - \underbrace{(\bar{w}^{*T} \cdot \bar{\bar{X}}^T \cdot F)} - \underbrace{(F^T \cdot \bar{\bar{X}} \cdot w)} + \bar{w}^T (\bar{\bar{X}}^T \bar{\bar{X}}) \bar{w}^*$$

are the same

Take gradient and set to zero.

$$\boxed{(\bar{\bar{X}}^T \cdot \bar{\bar{X}}) \cdot \bar{w}^*_{OPT} = \bar{\bar{X}}^T \cdot F} \longrightarrow \text{Normal equations.}$$

- check they hold for $\boxed{k=1}$ by explicitly plugging them.

$$\boxed{\bar{w}^*_{OPT} = (\bar{\bar{X}}^T \bar{\bar{X}})^{-1} \cdot \bar{\bar{X}}^T \cdot F}$$

MLE interpretation :-

Suppose data is generated as :-

$$f(\bar{x}) = w_0 + w_1 \bar{x}_1 + \cdots + \bar{w}_k \cdot \bar{x}_k + \underbrace{\varepsilon}_{N(0,\sigma^2)}$$

gaussian.

$$P(f_1, \ldots, f_n \mid \bar{\bar{X}}, \bar{w}^*, \sigma^2)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(f_i - \bar{x}_i^T \cdot \bar{w}^*)^2}{2\sigma^2} \right)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \cdot \exp\left(- \frac{1}{2\sigma^2} \cdot \|F - \bar{\bar{X}} \cdot \bar{w}^*\|_2^2 \right)$$

Gradient descent now