Lecture-10 (Logistic Regression)

Probabolistic :- $P(Y=y \mid X=x) = \dfrac{1}{1+\exp(y\langle\theta,x\rangle)}$

$y = \boxed{+1}$ or $\boxed{-1}$ $= \dfrac{1}{1+\exp(\langle\theta,x\rangle)}$ $\dfrac{1}{1+\exp(-1\langle\theta,x\rangle)}$

→ $P(Y=1/x) + p(Y=-1/x=1) = 1$

→ Decision boundary $p(Y=1/x) = p(Y=-1/x) = 0.5$

→ $p(Y=1/x) > 0.5$ ⇒ greater Confidence lasel 1 than -1

Prediction Confidence

Task	Use
→ predict the label associted with a feature vector x	→ predict rule $\text{Sign}(\theta, x)$
→ Measure of Confidence of that prediction	→ $P(Y=y \mid X=x) = \dfrac{1}{1+\exp(y\langle\theta,x\rangle)}$

Properties of Linear classifiers

↳ (1) Easy to train

↳ (2) Can predict labels very fast

(If x is sparse, the inner product $\langle\theta,x\rangle$ is Extremely fast)

↳ (3) Statistical theory of linear classifiers leading to Effective modeling strategies.

(4) Linear classifers - Excel at higher dimensions

↳ attractive computation load

Bias Term $\langle\theta,x\rangle + c$

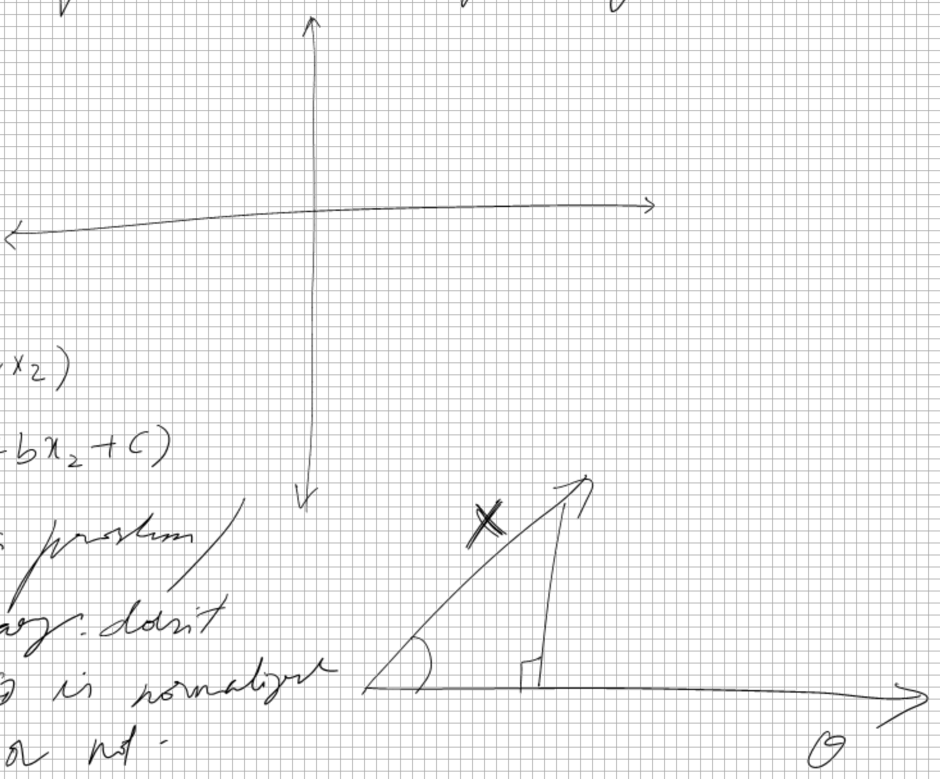
$= x_1\theta_1 + x_2\theta_2 + \cdots + x_d\theta_d + \boxed{c}$

Bias term makes our model

[illegible] necessarily not pass through origin.

→ Linear classifiers have the form $y = \text{sign}(\langle \theta, x \rangle)$

→



→ Given $x = (x_1, x_2)$

→ $\text{sign}(ax_1 + bx_2 + c)$

→ classification problem/

decision boundary doesn't affected whether θ is normalized or not.

→ Should the decision boundary pass through origin?

⇒ X or parameter vector is zero.

Answer:- Not Necessarily.

If we donot require that (by assigning our features to value 1) the classifier becomes considerably more powerful.

i.e we can learn much more general decision boundaries

$$x_1\theta_1 + x_2\theta_2 + - - + x_d\theta_d + c$$

without offset, the decision boundary has to to pass through Origin. i.e ⇒ $\langle \theta, x \rangle = 0$

at origin $x = 0, \theta = 0 \Rightarrow \langle \theta, x \rangle = 0$

→ Decision boundary passing through origin restricts the model in a significant way.

$$x_1\theta_1 + x_2\theta_2 + \cdots + x_d\theta_d + c = \underline{\langle x,\theta\rangle + c}$$

$$\langle X, \theta \rangle$$

↳ θ is d+1 dimensions.

$$(x_1, x_2 \cdots x_d, \underline{\underline{1}})$$

→ **MLE**

Frequentists:- MLE uses pairs of feature vectors & label, usually from historic data to Estimate correct classifier or nature.

Bayesian:- A single classifier cannot represent the "truth". Estimate the desired probability that each classifier is correct & use them all.

Justification for using MLE:

⇒ It converges to the optimal solutions in the limit of large data (consistency).

⇒ Convergence occurs at the fastest possible rate of convergence (statistical Efficiency).

Gradient Descent for Logistic:- is a scalable to large data as long it is sparse, not so much dependent on the dimensionality.

dot product $\langle \theta, x \rangle$ = easy to calculate

$$\theta_j \longleftarrow \theta_j - \alpha \frac{\partial \sum_{i=1}^{n} \log\left(1 + \exp\left(y^{(i)} \langle \theta, x^{(i)} \rangle\right)\right)}{\partial \theta_j}$$

when $x^{(i)}$ are sparse, partial derivative can be made particularly fast.

→ Stochastic Gradient converges faster than Gradient descent.