Given a sample x, find its label y.

$$y' = \arg\max_y p(y|x). \leftarrow \text{Discriminative model.}$$

$$= \arg\max_y \frac{p(x|y)\cdot p(y)}{p(x)}$$

$$= \arg\max_y p(x|y)\cdot p(y). \leftarrow \text{Genative model.}$$

- Discriminative Model $P(y|x;\theta)$

$$\begin{cases} \text{Regression: Linear Regression: } y|x \sim N(h(\theta), \sigma^2) \\ \text{Binary Classification: Logistic Regression: } y|x \sim \text{Bern}(h_\theta(x)) \\ \text{Multi-classification: Softmax Regression: } y|x \sim \text{Multi}(h_\theta(x)). \end{cases}$$

▷ LR : Given x, we have: $y = (\theta^*)^T x + b^*$

▷ Classification: Given x, $h_\theta(x) = \begin{bmatrix} p(y=1|x) \\ \vdots \\ p(y=k|x) \end{bmatrix}$.

e.g. $h_\theta(x) = \begin{array}{c}\begin{bmatrix} 0.1 \\ 0.2 \\ 0.7 \end{bmatrix}\end{array}\begin{array}{l} y=0 \\ y=1 \\ \leftarrow y=2 \end{array}$, $y' = \arg\max_y P(Y=y|x)$.

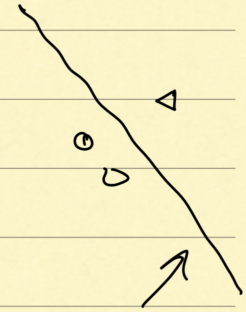
- Generative Model: $p(x,y) = \underline{p(x|y)}\cdot p(y)$.

$$\begin{cases} \text{Continuous } \overset{x}{\text{Input}}\text{: GDA} & \begin{cases} y \sim \text{Bern}(\phi) \\ x|y=b \sim N(\mu_b, \Sigma). \end{cases} \\ \text{Discrete Input: NB} & \begin{cases} p(x_1,\dots,x_n|y) = \prod_{i=1}^{n} p(x_i|y). \\ y \sim \text{Bern}(\phi) \\ x_i|y=b \sim \text{Bern}(\phi_{i|y=b}). \end{cases} \end{cases}$$

- ↑ Statistical View Point.
  ↓ Optimization.

  we want a standard / Criteria
  to tell me $y = -1, 1$.

  { Draw a line
  { After transform, to draw a line.



Details :

① Linear Regression :

- $h(x;\theta,b) = \theta^T x + b$, $\theta, x \in \mathbb{R}^n$, $b, y \in \mathbb{R}$.
- $J(\theta) = \frac{1}{2} \sum_{i=1}^{m} (h(x^{(i)}) - y^{(i)})^2$

  $= \frac{1}{2} (X\theta - y)^T (X\theta - y)$.

- Normal Equation : $\nabla_\theta J(\theta) = X^T X \theta - X^T y$

  $\theta^* = (X^T X)^{-1} \cdot X^T y$.

  $(X^T X + \alpha I)^{-1} \cdot X^T y$. : Ridge Reg

- Gradient Descent : $\theta' := \theta - \alpha \cdot \nabla J(\theta)$.

  $\theta_1 = \theta_0 - \alpha \cdot \nabla J(\theta_0)$.

  $\theta_2 = \theta_1 - \alpha \cdot \nabla J(\theta_1)$

  $= \underset{\downarrow\, 0.}{\theta_0} - \alpha \cdot \nabla J(\theta_0) - \alpha \cdot \nabla J(\theta_1)$.

- Newton's Method.

- Relationship with MLE
  - MLE : maximize $\prod_{i=1}^{m} P(y^{(i)} | x^{(i)})$, log-likelihood.
  - Assumption : $y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)}$

    where $\varepsilon^{(i)} \underset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$.

    $\downarrow$

    $y|x$.
  - LR = MLE + Linear Model + Gaussian Assumption.

② Logistic Regression : $y = 0, 1$.

- $y|x \sim \mathrm{Bern}(h_\theta(x))$. ↱ hypothesis function.

  i.e. $P(y=1|x) = h_\theta(x) = (h_\theta(x))^1 (1-h_\theta(x))^{1-1}$

  where $h_\theta(x) \triangleq \sigma(\theta^T x) \triangleq \dfrac{1}{1+e^{-\theta^T x}}$ ☆

  ↑ sigmoid function

③ Softmax Regression : $y = 0, 1, \cdots, k-1$.

- $y|x \sim \mathrm{Multinomial}(h_\theta(x))$.

  where $h_\theta(x) = \begin{bmatrix} P(y=0|x) \\ \vdots \\ P(y=k-1|x) \end{bmatrix} = \dfrac{1}{\sum_{j=0}^{k-1} e^{\theta_j^T x}} \begin{bmatrix} e^{\theta_0^T x} \\ \vdots \\ e^{\theta_{k-1}^T x} \end{bmatrix}$

  ↑ normalize term, ↑ $1 > h_\theta(x) > 0$

$v \succ 1$ $\quad v(i) > 1$

$\begin{bmatrix} v(1) \\ \vdots \\ v(k) \end{bmatrix} \succ 1$

- $l(\Theta) = \sum_{i=1}^{m} \log P(y^{(i)} | x^{(i)}).$

$$= \sum_{i=1}^{m} \log \prod_{l=0}^{k-1} P(y^{(i)} = l | x^{(i)})^{1\{y^{(i)}=l\}}.$$

$$= \sum_{i=1}^{m} \sum_{l=0}^{k-1} 1\{y^{(i)} = l\} \cdot \log \frac{e^{\theta_l^T x^{(i)}}}{\sum_{j=0}^{k-1} e^{\theta_j^T x^{(i)}}}$$

- $\nabla_{\theta_1}(l(\Theta)) = \nabla_{\theta_1} \sum_{i=1}^{m} \sum_{l=0}^{k-1} 1\{y^{(i)} = l\} \cdot \log \frac{e^{\theta_l^T x^{(i)}}}{\sum_{j=0}^{k-1} e^{\theta_j^T x^{(i)}}}$

$$= \sum_{i=1}^{m} \nabla_{\theta_1} \sum_{l=0}^{k-1} 1\{y^{(i)} = l\} \left(\theta_l^T x^{(i)} - \log \sum_{j=0}^{k-1} e^{\theta_j^T x^{(i)}} \right).$$

$$= \sum_{i=1}^{m} \nabla_{\theta_1} \left(\sum_{l=0}^{k-1} 1\{y^{(i)} = l\} \cdot \theta_l^T \cdot x^{(i)} - \sum_{l=0}^{k-1} 1\{y^{(i)} = l\} \log \sum_{j=0}^{k-1} e^{\theta_j^T x^{(i)}} \right).$$

$$= \sum_{i=1}^{m} \nabla_{\theta_1} \left(\sum_{l=0}^{k-1} 1\{y^{(i)} = l\} \cdot \theta_l^T \cdot x^{(i)} - \log \sum_{j=0}^{k-1} e^{\theta_j^T x^{(i)}} \right).$$

$=$

$$\nabla_{\theta_1} \sum_{l=0}^{k-1} 1\{y^{(i)} = l\} \cdot \theta_l^T \cdot x^{(i)}$$

$$= \nabla_{\theta_1} \left(\underline{1\{y^{(i)} = 0\} \cdot \theta_0^T \cdot x^{(i)}} + \dots + \cancel{1\{y^{(i)} = k-1\} \cdot \theta_{k-1}^T \cdot x^{(i)}} \right).$$

$0.$

$$= \nabla_{\theta_1} \cdot 1\{y^{(i)} = 1\} \cdot \theta_1^T \cdot x^{(i)}$$

$$= 1 \cdot \{y^{(i)} = 1\} \cdot x^{(i)}$$

- Relation with Logistic Regression.

$k = 2.$

$$P_{Y|X}(1|x) = \frac{e^{\theta_1^T x}}{e^{\theta_0^T x} + e^{\theta_1^T x}} = \frac{1}{e^{(\theta_0 - \theta_1)^T x} + 1} = \sigma((\theta_1 - \theta_0)^T x).$$

④ Gaussian Discriminative Analysis (GDA) / QDA.

- $p(x,y) = p(x|y)\cdot p(y)$.

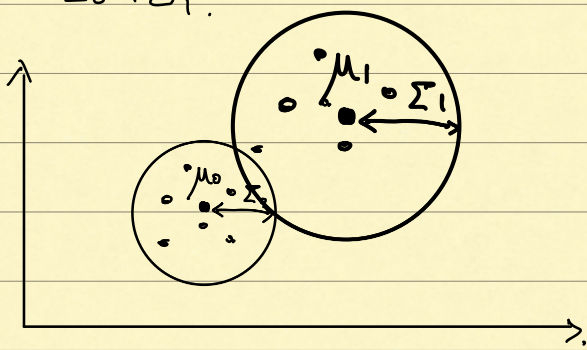
Assume $y \sim \mathrm{Bern}(\phi)$, $\phi \in \mathbb{R}$

$x|y=0 \sim \mathcal{N}(\mu_0, \Sigma_0)$, $\mu_0, \mu_1 \in \mathbb{R}^n$

$x|y=1 \sim \mathcal{N}(\mu_1, \Sigma_1)$. , $\Sigma_0, \Sigma_1 \in \mathbb{R}^{n\times n}$.

- GDA: $\Sigma_0 = \Sigma_1$

QDA: $\Sigma_0 \neq \Sigma_1$.



Intuitions:

① Y

Y	0	1
p(y)	$\frac{1}{10}$	$\frac{9}{10}$

ϕ: how many labeled samples

② are drawn Datasets.

μ_1: Q_1: the mean of samples which are labeled 1.

Σ_1: $(x^{(i)} - \mu_1)(x^{(i)} - \mu_1)^T$.

①
$$\phi = \frac{\sum_{i=1}^{m} 1\{y^{(i)}=1\}}{m}$$

$$\mu_1 = \frac{\sum_{i=1}^{m} 1\{y^{(i)}=1\}\cdot x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)}=1\}}$$

$$\Sigma_1 = \frac{\sum_{i=1}^{m} 1\{y^{(i)}=1\}\cdot(x^{(i)}-\mu_1)(x^{(i)}-\mu_1)^T}{\sum_{i=1}^{m} 1\{y^{(i)}=1\}}.$$

- $$\ell(\phi, \mu_0, \mu_1, \Sigma_0, \Sigma_1) \triangleq \log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)})$$
$$= \sum_{i=1}^{m} \log p(x^{(i)}, y^{(i)})$$
$$= \sum_{i=1}^{m} \left(\log p(x^{(i)} | y^{(i)}) + \log p(y^{(i)}) \right).$$

① Do one thing in one line.

- $y \sim \text{Bern}(\phi) \rightarrow p(y) = \phi^{y}(1-\phi)^{1-y}$, $y = 0, 1$.
$$\log p(y) = y \log \phi + (1-y) \log(1-\phi).$$
$$x|y=0 \sim \mathcal{N}(\mu_0, \Sigma_0) \rightarrow p(x|y=0) = \left[(2\pi)^{\frac{n}{2}} \cdot |\Sigma_0|^{\frac{1}{2}}\right]^{-1} \cdot \exp\left(-\frac{1}{2}(x-\mu_0)^T \Sigma_0^{-1} (x-\mu_0)\right).$$
$$\log p(x|y=0) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_0| - \frac{1}{2}(x-\mu_0)^T \cdot \Sigma_0^{-1} (x-\mu_0).$$

$$x|y=b \sim \mathcal{N}(\mu_b, \Sigma_b) \rightarrow \log p(x|y=b) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_b| - \frac{1}{2}(x-\mu_b)^T \cdot \Sigma_b^{-1} (x-\mu_b).$$

where $b = 0, 1$.

② If you want to derive the equation for all cases, try to figure out a special case.

e.g. WA.1.4. $\nabla_{\theta_t} \ell(\Theta) \rightarrow \nabla_{\theta_1} \ell(\Theta)$.

- $$\ell = \sum_{i=1}^{m} \left(\log p(x^{(i)} | y^{(i)}) + \log p(y^{(i)}) \right)$$
$$= \sum_{i=1}^{m} -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_{y^{(i)}}| - \frac{1}{2}(x^{(i)} - \mu_{y^{(i)}})^T \cdot \Sigma_{y^{(i)}}^{-1} (x^{(i)} - \mu_{y^{(i)}})$$
$$+ \sum_{i=1}^{m} y^{(i)} \log \phi + (1-y^{(i)}) \log(1-\phi).$$
$$= \sum_{i=1}^{m} \sum_{b=0}^{1} 1\{y^{(i)} = b\} \left[-\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_b| - \frac{1}{2}(x^{(i)} - \mu_b)^T \cdot \Sigma_b^{-1} (x^{(i)} - \mu_b) \right]$$
$$+ \sum_{i=1}^{m} y^{(i)} \log \phi + (1-y^{(i)}) \log(1-\phi)$$

- $\frac{\partial l}{\partial \phi} = 0, \quad \frac{\partial l}{\partial \mu_b} = 0, \quad \frac{\partial l}{\partial \Sigma_b} = 0, \quad b = 0, 1$

- $$\frac{\partial l}{\partial \phi} = \frac{\partial}{\partial \phi} \sum_{i=1}^{m} y^{(i)} \log \phi + (1 - y^{(i)}) \log (1 - \phi)$$
$$= \sum_{i=1}^{m} y^{(i)} \cdot \frac{1}{\phi} + (1 - y^{(i)}) \frac{-1}{1 - \phi} = 0$$

$$\Rightarrow \phi^* = \frac{1}{m} \sum_{i=1}^{m} 1\{y^{(i)} = 1\}.$$

- $$\frac{\partial l}{\partial \mu_0} = \frac{\partial}{\partial \mu_0} \sum_{i=1}^{m} \sum_{b=0}^{1} 1\{y^{(i)} = b\} \left[-\frac{1}{2} (x^{(i)} - \mu_b)^T \cdot \Sigma_b^{-1} (x^{(i)} - \mu_b) \right]$$
$$= \frac{\partial}{\partial \mu_0} \sum_{i=1}^{m} 1\{y^{(i)} = 0\} \left[-\frac{1}{2} (x^{(i)} - \mu_0)^T \cdot \Sigma_0^{-1} (x^{(i)} - \mu_0) \right]$$
$$+ \frac{\partial}{\partial \mu_0} \sum_{i=1}^{m} 1\{\underline{y^{(i)}} = 1\} \left[-\frac{1}{2} (x^{(i)} - \mu_1)^T \cdot \Sigma_1^{-1} (x^{(i)} - \mu_1) \right]$$
$$= \frac{\partial}{\partial \mu_0} \sum_{i=1}^{m} 1\{y^{(i)} = 0\} \left(-\frac{1}{2} (x^{(i)} - \mu_0)^T \cdot \Sigma_0^{-1} (x^{(i)} - \mu_0) \right)$$
$$= \cancel{\frac{\partial}{\partial \mu_0}} \sum_{i=1}^{m} 1\{y^{(i)} = 0\} \cdot \frac{\partial}{\partial \mu_0} \left[-\frac{1}{2} (x^{(i)} - \mu_0)^T \cdot \Sigma_0^{-1} (x^{(i)} - \mu_0) \right] \quad (*)$$

Recall: $\frac{\partial}{\partial v} (v - w)^T A (v - w) = (A + A^T)(v - w).$

$$(*) = \sum_{i=1}^{m} 1\{y^{(i)} = 0\} \cdot \left[-\Sigma_0^{-1} (\mu_0 - x^{(i)}) \right].$$
$$= -\Sigma_0^{-1} \cdot \left(\sum_{i=1}^{m} 1\{y^{(i)} = 0\} \cdot (\mu_0 - x^{(i)}) \right) = 0.$$
$$\Rightarrow \sum_{i=1}^{m} 1\{y^{(i)} = 0\} \cdot (\mu_0 - x^{(i)}) = 0$$
$$\Rightarrow \mu_0^* = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 0\} \cdot x^{(i)}}{\sum_{j=1}^{m} 1\{y^{(j)} = 0\}}$$

$$\bullet\ \frac{\partial l}{\partial \Sigma_0} = \frac{\partial}{\partial \Sigma_0} \cdot \sum_{i=1}^{m} \sum_{b=0}^{1} 1\{y^{(i)}=b\} \left(-\frac{1}{2}\log|\Sigma_b| - \frac{1}{2}(x^{(i)}-\mu_b)^T \cdot \Sigma_b^{-1} (x^{(i)}-\mu_b)\right)$$

$$= \sum_{i=1}^{m} 1\{y^{(i)}=0\} \cdot \frac{\partial}{\partial \Sigma_0} \left[-\frac{1}{2}\log|\Sigma_0| - \frac{1}{2}(x^{(i)}-\mu_0)^T \cdot \Sigma_0^{-1} \cdot (x^{(i)}-\mu_0)\right] \quad (**)$$

Note that $\bullet\ \frac{\partial}{\partial \Sigma}\log|\Sigma| = \frac{1}{|\Sigma|} \cdot \frac{\partial |\Sigma|}{\partial \Sigma} = \frac{1}{|\Sigma|} \cdot |\Sigma| \cdot \Sigma^{-1} = \Sigma^{-1}$

$\bullet\ \frac{\partial}{\partial \Sigma} v^T \Sigma^{-1} \Sigma = -\Sigma^{-1} \cdot v v^T \cdot (\Sigma^{-1})^T$.

$$(**) \Leftrightarrow = \sum_{i=1}^{m} 1\{y^{(i)}=0\} \cdot \left[-\frac{1}{2} \cdot \Sigma_0^{-1} + \frac{1}{2}\Sigma_0^{-1} \cdot (x^{(i)}-\mu_0)(x^{(i)}-\mu_0)^T \cdot \Sigma_0^{-1}\right] = 0.$$

$$\Sigma_0^* = \frac{\sum_{i=1}^{m} 1\{y^{(i)}=0\} \cdot (x^{(i)}-\mu_0)(x^{(i)}-\mu_0)^T}{\sum_{i=1}^{m} 1\{y^{(i)}=0\}}.$$

1. Treat notation clearly
2. One thing at One Step.
3. Too many notations? Try a special case.
4. Burden will ↓, when the diff are known.
5. Turn your intuition radar on?
   e.g scalar/ vector? Make sense?

- Relation to Logistic Regression. ☆. Lec4. P23/28.
  - $p(y=1|x) = \frac{1}{1+e^{-\theta^T x}}$
  - If $x|y \sim \mathcal{N}(\mu, \Sigma)$, $p(y|x)$ is a logstic function.

⑤ Naive Bayes.

- Dictionary

$$x = \left.\begin{array}{c} \left[\begin{array}{c} 0 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 0 \end{array}\right] \end{array}\begin{array}{l} a \\ \\ \text{Review} \\ \text{Session} \\ \\ \\ \end{array}\right\} \text{large.}$$

spam? not spam?

$$p(x_1, \cdots, x_n | y) = p(x_1|y) \cdots p(x_n|y)$$
$$= \prod_{i=1}^{n} p(x_i|y)$$

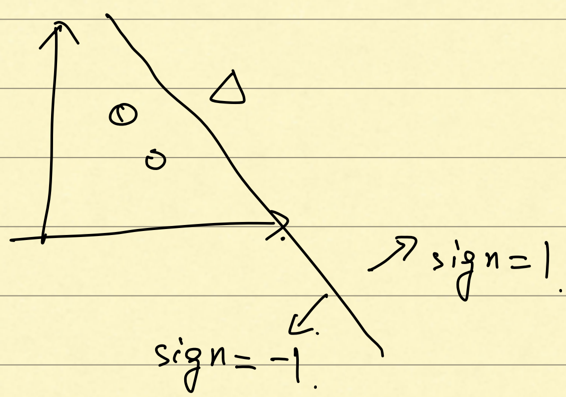
- $y \sim \text{Bern}(\phi)$

$x_i | y=b \sim \text{Bern}(\phi_{i|y=b})$, $b = 1, 2.$

- $$\phi^* = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}{m}$$

$$\phi^*_{j|y=1} = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\} \cdot 1\{x_j^{(i)} = 1\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}$$

⑥ SVM.



- Given $\{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$, find (w^*, b^*) to.

prime problem: $\min\limits_{w,b} \frac{1}{2}\|w\|^2$

$\quad$ s.t $y^{(i)}(w^T x^{(i)} + b) \geq 1$, $i = 1, \cdots, m$.

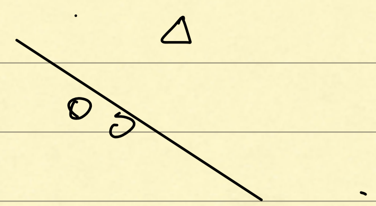
dual problem: $\max\limits_{\alpha} \sum\limits_{i=1}^{m} \alpha_i - \frac{1}{2} \sum\limits_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} \langle x^{(i)}, x^{(j)} \rangle$

$\quad$ s.t. $\alpha_i \geq 0$, $i = 1, \cdots, m$

$\quad\quad$ $\sum \alpha_i y^{(i)} = 0$.

$\Rightarrow \alpha^*$

Solution:
$$\begin{cases} w^* = \sum\limits_i \alpha_i^* y^{(i)} \cdot x^{(i)} \\ b^* = -\frac{1}{2}\left(\max\limits_{i: y^{(i)} = -1} w^{*T} x^{(i)} + \min\limits_{i: y^{(i)} = 1} (w^*)^T x^{(i)} \right) \end{cases}$$



Given new sample z

$$y = \operatorname{sign}[w^{*T} z + b^*] = \operatorname{sign}\left[\sum_{i=1}^{m} \alpha_i y^{(i)} \langle x^{(i)}, z \rangle + b^*\right].$$

where $\operatorname{sign}(t) = \begin{cases} 1, & t > 0 \\ 0, & t = 0 \\ -1, & t < 0. \end{cases}$

- Soft-SVM $\quad$ Review at home ☆.

• Kernel Trick.

Messy Data.

$\phi: \underset{\mathbb{R}^d}{\underset{\Uparrow}{x}} \to \underset{\mathbb{R}^D}{\underset{\Uparrow}{\phi(x)}}$.

$K(x, x') \triangleq \phi^T(x)\phi(x') \in \mathbb{R}$. $\quad K \in \mathbb{R}^{m \times m}$.

dual problem: $\max\limits_{\alpha} \sum\limits_{i=1}^{m} \alpha_i - \frac{1}{2} \sum\limits_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} \underline{K(x^{(i)}, x^{(j)})}$

s.t. $C \geq \alpha_i \geq 0, \quad i = 1, \cdots, m$

$\sum \alpha_i y^{(i)} = 0.$

$\Rightarrow \alpha^*$

Solution: $\begin{cases} w^* = \sum\limits_i \alpha_i^* y^{(i)} \phi(x^{(i)}) \\ b^* = y^{(j)} - \sum\limits_{i=1}^{m} \alpha_i^* y^{(i)} K(x^{(i)}, x^{(j)}) \text{ for some } j. \end{cases}$

• $f(x) = w^{*T}\phi(x) + b^*$

$= \sum\limits_{i=1}^{m} \alpha_i y^{(i)} \langle \phi(x^{(i)}), \phi(x) \rangle + b.$

• $y' = \arg\max\limits_{y} P(y|x)$

$= \arg\max\limits_{y} P(x|y) \cdot P(x).$

$y = \text{sign}(\_\_)$