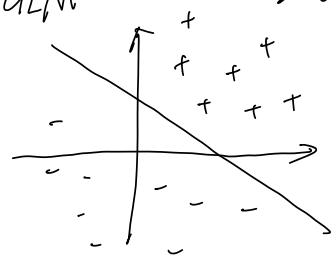
GLM



Binary classification

$f: X \to Y$

$\mathbb{R}^2 \quad \mathbb{R}$.

$Y^+ = \{y \mid y > \text{threshold}\}$

$Y^- = \{y \mid y \le \text{threshold}\}$

$P(y \in Y^+ \mid x) = \phi$

$P(y \in Y^- \mid x) = 1-\phi$

$(y \mid x; \phi) \sim \text{Bernoulli}(\phi)$

Exponential Family.

$P(y;\eta) = b(y)\, e^{\eta^T T(y) - a(\eta)}$.

property 1:

$e^x \log x \Rightarrow x = e^{\log x}$

$P(y;\phi) = \phi^y (1-\phi)^{1-y} = e^{(\log \phi^y (1-\phi)^{1-y})}$

$= e^{y\log\phi + (1-y)\log(1-\phi)}$

$= e^{y[\log\phi - \log(1-\phi)] + \log(1-\phi)}$.

Bernoulli distribution $= e^{y\log\frac{\phi}{1-\phi} - (-\log(1-\phi))}$

$\boxed{\eta = \log\frac{\phi}{1-\phi}}$

$b(y) = 1$

$T(y) = y$

$a(\eta) = -\log(1-\phi)$.

① canonical link Function $\quad g(\eta) = \mathbb{E}[T(y);\eta]$

② canonical response Function $\quad \eta = g^{-1}(\mathbb{E}[T(y)|\eta])$

Generalized linear model

1. $y|x;\theta \sim$ Exponential Family
2. $h(x) = \mathbb{E}[T(y)|x]$
3. $\eta$ varies linearly with $x$

$\eta = \theta^T x$

1. $y|x;\phi \sim$ Bernoulli$(\phi)$
2. $T(y) = y \quad h(x) = \mathbb{E}[y|x] = \phi$
3. $\eta = \theta^T x$

$h(x) = \frac{1}{1+e^{-\theta^T x}}$

$= \text{sigmoid}(\eta)$

$= \frac{1}{1+e^{-\eta}}$

$\{(x_i, y_i)\}_{i=1}^{m}$ → Label.

↑ complex world

$y' = \arg\max_y p(y|x)$. → Discriminative Model.

$= \arg\max_y p(x|y) \cdot p(y)$ → Generative Model.

↳.

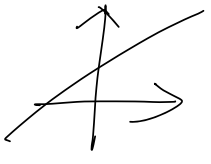
Basic knowledge.

- Discriminative Model: $p(y|x;\theta)$.
  - Regression: Linear Regression $y|x \sim \mathcal{N}(h_\theta(x), \sigma^2)$.
  - Binary class...: Logistic Regression $y|x \sim Bern(h_\theta(x))$
  - Multi-class ...: Softmax ... $y|x \sim$ Multinomial$(h_\theta(x))$.

▷ Linear model: Given $x$, $y = \theta^T x + b$

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

$= \begin{bmatrix} 0.1 \\ 0.2 \\ 0.7 \end{bmatrix}$ ←



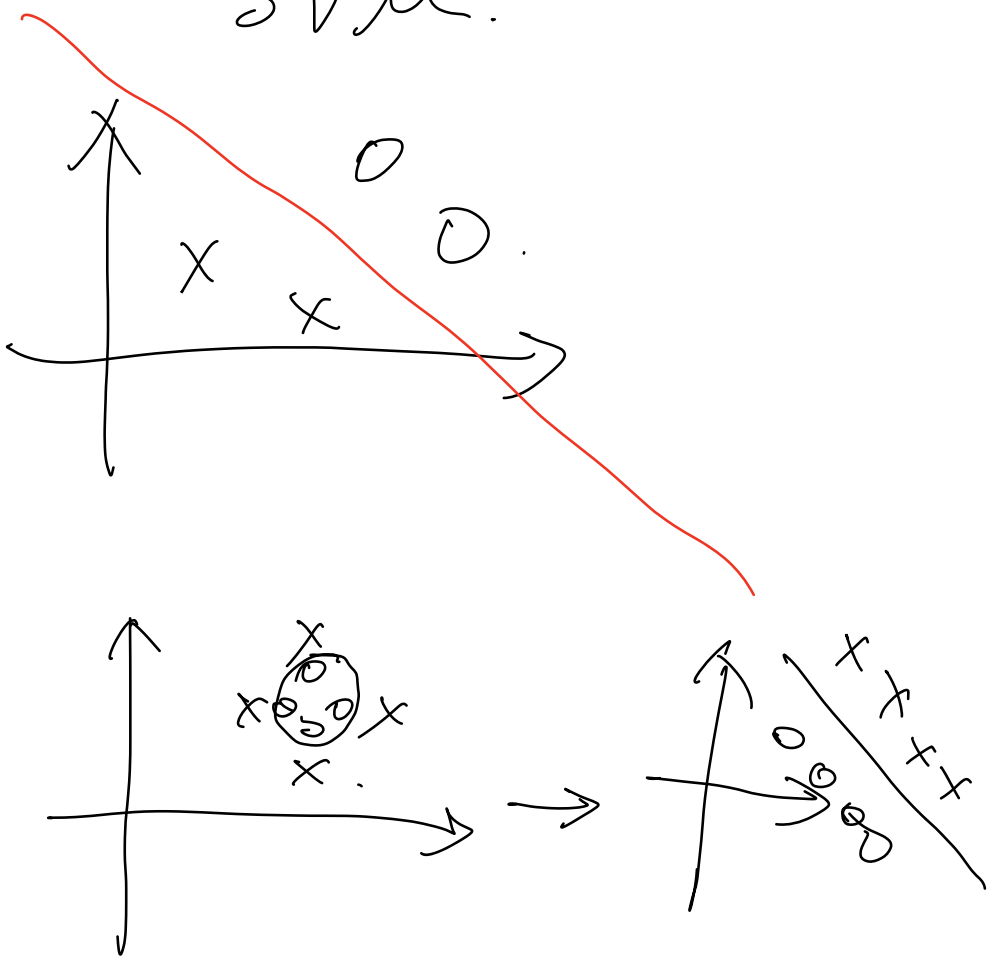
- Generative Model: $p(x,y) = p(x|y) \cdot p(y)$.
  - Continuous Input: GDA
    - $y \sim Bern(\phi)$.
    - $x|y=b \sim \mathcal{N}(\mu_b, \Sigma)$.
  - Discrete Input: NB.
    - $p(x_1, \dots, x_n|y) = \prod_{i=1}^{n} p(x_i|y)$.
    - $y \sim Bern(\phi)$.
    - $x_i|y=b \sim Bern(\phi_{i|y=b})$.

▷ Bayes Rule.

↑ statistical

↓ Optimization.

SVM.



① Linear Regression.

$$y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)}$$

$\varepsilon^{(i)} \leftarrow \mathcal{N}(0, \sigma^2)$.

LR = MLE + Linear Model + Gaussian Ass

Laplacian.

18:41

② $\theta \sim N(0, \frac{1}{2\lambda} I)$. MAP.

③ Softmax function with

$k = 2$

$$P_{Y|X}(1|x) = \frac{e^{\theta_1^T x}}{e^{\theta_1^T x} + e^{\theta_2^T x}} = \frac{1}{1 + e^{-(\theta_1 - \theta_2)^T x}} = \sigma((\theta_1 - \theta_2)^T x)$$

Sigmoid.

$$\theta = \arg\max_\theta p(\theta | \{\ \})$$

$$= \arg\max_\theta p(\{x, y\} | \theta) \cdot p(\theta)$$

Gaussian Discriminative Analysis (GDA) → QDA

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

$x|y=0 \sim N(\mu_0, \Sigma_0)$

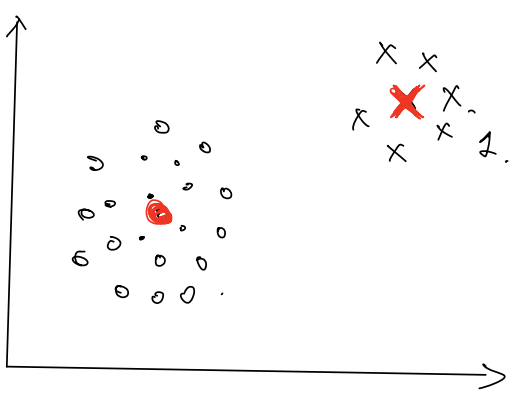
$x|y=1 \sim N(\mu_1, \Sigma_1)$

$y \in \{0, 1\}$.

$(x_1, y_1)$

$\vdots$

$(x_m, y_m)$.





1.

$$\phi = \frac{\sum_{i=1}^m 1\{y_i = 0\}}{m}$$

$$\mu_b = \frac{\sum_{i=1}^m 1\{y_i = b\} \cdot x_i}{\sum_{i=1}^m 1\{y_i = b\}}, \quad b = 0, 1.$$

$$\Sigma_b = \frac{\sum_{i=1}^m 1\{y_i = b\} (x^{(i)} - \mu_b)(x^{(i)} - \mu_b)^T}{\sum_{i=1}^m 1\{y_i = b\}}$$

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

- $y \sim \text{Bern}(\phi) \Rightarrow \quad p(y) = \phi^y (1-\phi)^{1-y}, \quad y = 0, 1.$
$$= \begin{cases} \phi^0 \cdot (1-\phi)^1, & y = 0. \\ \phi^1 (1-\phi)^0, & y = 1 \end{cases}$$
$$\log p(y) = y \cdot \log \phi + (1-y) \log (1-\phi). \leftarrow$$

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

$$x | y=1 \sim \mathcal{N}(\mu_1, \Sigma_1) \Rightarrow p(x|y=1) = \left( (2\pi)^{\frac{n}{2}} |\Sigma_1|^{\frac{1}{2}} \right)^{-1} \cdot \exp\left( -\frac{1}{2} (x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1) \right)$$
$$\log p(x|y=1) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_1|$$
$$- \frac{1}{2} (x-\mu_1)^T \cdot \Sigma_1^{-1} (x-\mu_1).$$

e.g. $x \in \mathbb{R}^{n+1}$, $\nabla_\theta \ell(\theta)$. $\nabla_{\theta_\ell}$

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

- $\frac{\partial \ell}{\partial \phi} = 0$. $\frac{\partial \ell}{\partial \mu_b} = 0$, $\frac{\partial \ell}{\partial \Sigma_b}$ , $\frac{\partial a^T x}{\partial x} = a$

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

- Relation to Logistic Regression. ☆
  - $p(y=1|x) = \frac{1}{1+e^{\theta^T x}}$

  If $p(x|y) \sim N(\mu, \Sigma)$. $p(y|x)$ is a logistic function.

Naive Bayes.

- Dictionary

$$x = \begin{bmatrix} 0 \\ 0 \\ \\ \\ 1 \\ 1 \end{bmatrix} \begin{matrix} a \\ abondon \\ \vdots \\ \\ Review \\ Session. \end{matrix} \Big\} \text{Large!!}$$

$x_1, x_2, \cdots, x_n$

$p(x_1, \cdots, x_n|y) = p(x_1|y) \cdots p(x_n|y)$

$= \prod_{i=1}^{n} p(x_i|y)$.

$y = \begin{cases} 0 & \text{no spam} \\ 1 & \text{spam!} \end{cases}$

- $y \sim Bern(\phi)$.

  $x_i|y=b \sim Bern(\phi_{i|y=b})$, $b=0,1$.

$$x^{(i)} = \begin{bmatrix} x_1^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix}.$$

$1 = 1\{y^{(i)}=0\} + 1\{y^{(i)}=1\}$.

$\ell(\phi_y, \phi_{j|y=0}, \phi_{j|y=1}, j=1,\cdots,n)$

$= \sum_{i=1}^{m} \log(p(x^{(i)}, y^{(i)}))$.

$= \sum_{i=1}^{m} \left( \log(p(x^{(i)}|y^{(i)})) + \log(p(y^{(i)})) \right)$

$= \sum_{i=1}^{m} \log \prod_{j=1}^{n} p(x_j^{(i)}|y^{(i)}) + \log p(y^{(i)})$

$= \sum_{i=1}^{m} \sum_{j=1}^{n} (1\{y^{(i)}=0\} + 1\{y^{(i)}=1\}) \log p(x_j^{(i)}|y^{(i)}) + \log p(y^{(i)})$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{b=0}^{1} 1\{y^{(i)}=b\}\cdot \log P(x_j^{(i)}\mid y^{(i)}) + \sum_{i=1}^{m} y^{(i)}\log(\phi_y) + (1-y^{(i)})\log(1-\phi_y)$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{b=0}^{1} 1\{y^{(i)}=b\}\left( x_j^{(i)}\log\phi_{j|y=b} + (1-x_j^{(i)})\log(1-\phi_{j|y=b})\right).$$

$$+ \sum_{i=1}^{m} y^{(i)}\log(\phi_y) + (1-y^{(i)})\log(1-\phi_y)$$

$$\frac{\partial \ell}{\partial \phi_y} = \frac{\partial}{\partial \phi_y}\left(\sum_{i=1}^{m} y^{(i)}\log(\phi_y) + (1-y^{(i)})\log(1-\phi_y)\right)$$

$$= \frac{\sum_{i=1}^{m} y^{(i)}}{\phi_y} - \frac{\sum_{i=1}^{m}(1-y^{(i)})}{1-\phi_y} = 0.$$

$$\frac{a}{b} = \frac{c}{d} = \frac{a+c}{b+d}.$$

$$\hookrightarrow \quad \phi_y = \frac{1}{m}\cdot 1\{y^{(i)}=y\}, \quad y=0,1.$$

Support Vector Machine. (SVM) (KKT)

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

prime problem

$$\min_{w,b} \frac{1}{2}\|w\|^2$$
$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i=1,\cdots,m.$$

dual problem:

$$\max_{\alpha} \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} \langle \alpha_i y^{(i)} x^{(i)}, \alpha_j y^{(j)} x^{(j)} \rangle$$
$$\text{s.t.} \quad \alpha_i \geq 0, \quad i=1,\cdots,m.$$
$$\sum_{i=1}^{m}\alpha_i y^{(i)} = 0.$$

Solution:

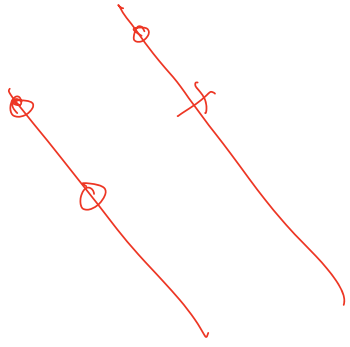
$$\begin{cases} w^* = \sum_i \alpha_i^* y^{(i)} x^{(i)} \\ b^* = -\frac{1}{2}\left(\max_{i:y^{(i)}=-1} w^{*T} x^{(i)} + \min_{i:y^{(i)}=1} w^{*T} x^{(i)}\right). \end{cases}$$

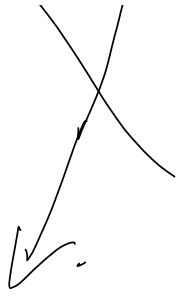
Given new sample $z$.

$$y = \text{sign}[w^{*T} z + b^*].$$

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

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



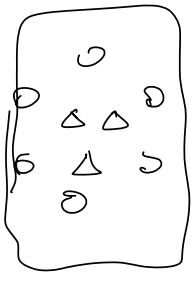


☆ Soft-SVM.

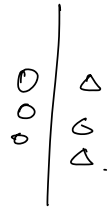
- kernel Trick.

$x \in \mathbb{R}^d \rightarrow \phi(x) \in \mathbb{R}^D$, $D \gg d$.

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



map



$\#\{x\} \times \#\{x\}$.

Dual: $\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j K(x^{(i)}, x^{(j)})$

s.t. $0 \le \alpha_i \le C$, $\sum \alpha_i y^{(i)} = 0$ $\Rightarrow \alpha^*$

Solution: $w^* = \sum \alpha_i^* y^{(i)} \phi(x^{(i)})$

$b^* = y^{(j)} - \sum_{i=1}^{m} \alpha_i^* y^{(i)} K(x^{(i)}, x^{(j)})$, for some $0 < \alpha_j < C$.

$f(w) = \text{sgn}(w^T \phi(x) + b)$.

$P_{XY}(x,y)$

$P_Y(y) = \sum_x P_{XY}(x,y)$

$P_X(x) = \sum_y P_{XY}(x,y)$

$P_{X|Y}$

$P_{Y|X}$