Foundations of Machine Learning (ECE 5984)

- Generative Learning Algorithms -

Eunbyung Park

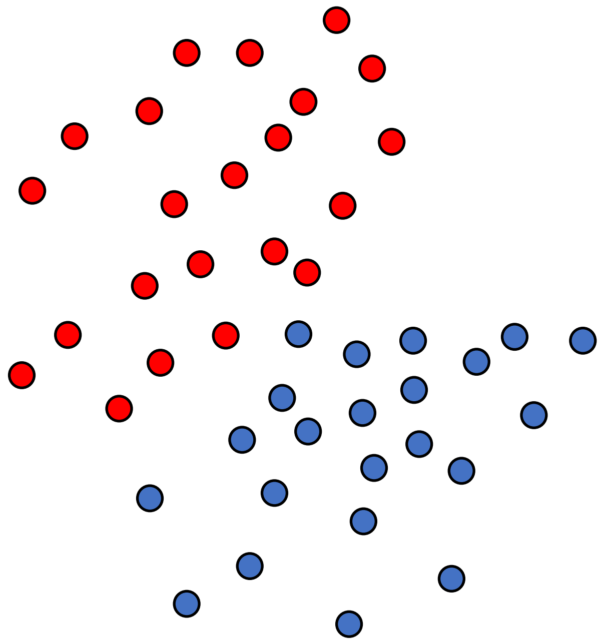
Assistant Professor

School of Electronic and Electrical Engineering

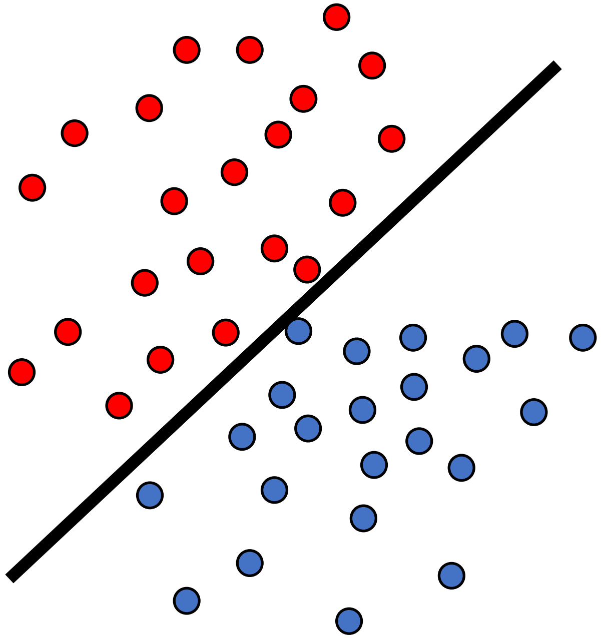
Eunbyung Park (silverbottlep.github.io)

Generative vs Discriminative

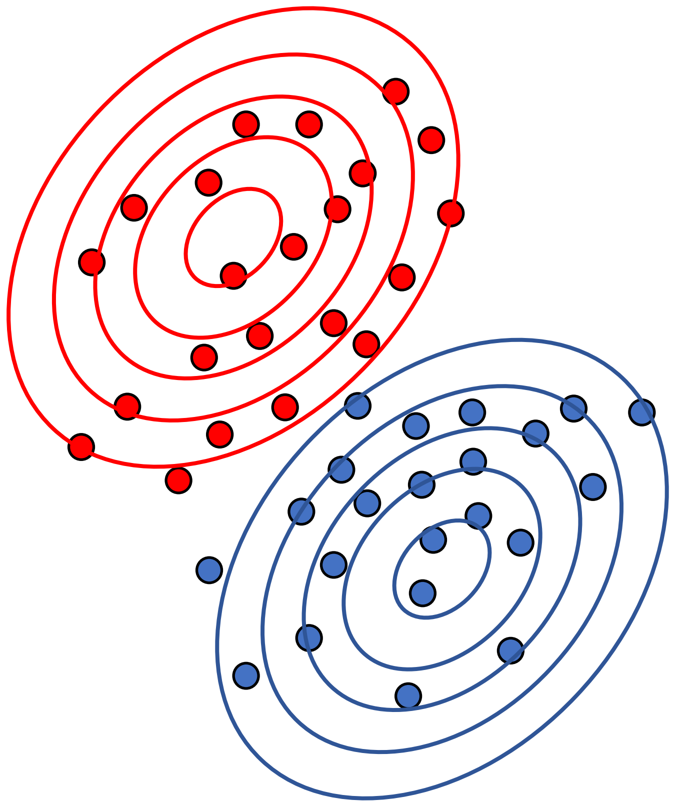
Discriminative Models



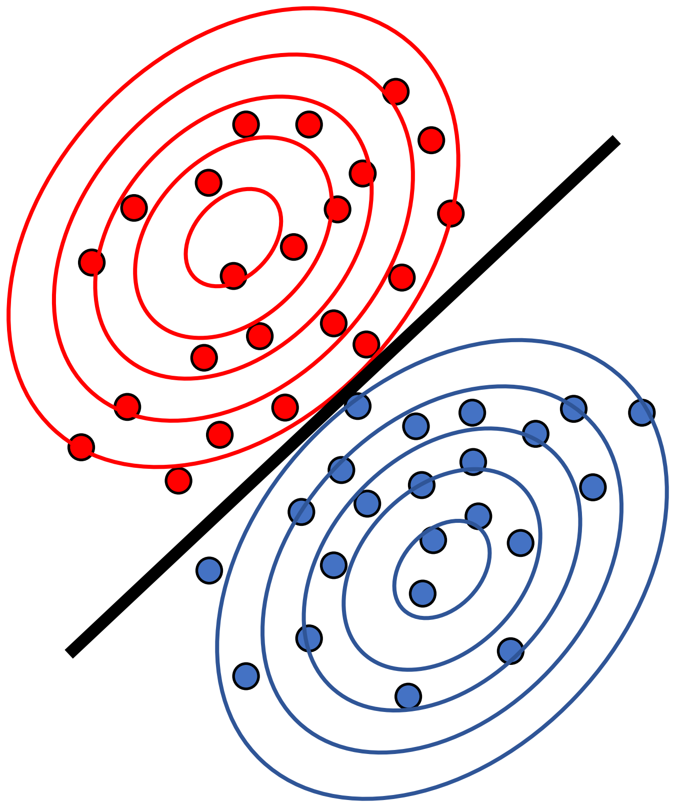
Discriminative Models



Generative Models



Generative Models



Discriminative vs. Generative

Discriminative Approach

$$p(y|x)$$

$$\underset{y}{\operatorname{argmax}}\, p(y|x)$$

Generative Approach

$$p(x|y), p(y)$$

$$\underset{y}{\operatorname{argmax}}\, p(y|x) = \underset{y}{\operatorname{argmax}}\, \frac{p(x|y)p(y)}{p(x)}$$

$$= \underset{y}{\operatorname{argmax}}\, p(x|y)p(y)$$

Gaussian Discriminant Analysis

Gaussian Discriminant Analysis

- Classification where input feature x are continuous variables

$$D = \{(x^{(1)}, y^{(1)}), \dots, (x^{(N)}, y^{(N)})\}, \qquad x^{(i)} \in \mathbb{R}^d, \qquad y^{(i)} \in \{0,1\}$$

Gaussian Discriminant Analysis

- Classification where input feature x are continuous variables

$$D = \{(x^{(1)}, y^{(1)}), \dots, (x^{(N)}, y^{(N)})\}, \qquad x^{(i)} \in \mathbb{R}^d, \qquad y^{(i)} \in \{0,1\}$$

$$\mu_0, \mu_1 \in \mathbb{R}^d \qquad \Sigma \in \mathbb{R}^{d \times d}$$

$$p(y) = \text{Bern}\,(\phi)$$
$$p(x|y=0) = N(\mu_0, \Sigma)$$
$$p(x|y=1) = N(\mu_1, \Sigma)$$

Shared Covariance

Gaussian Discriminant Analysis

- Classification where input feature x are continuous variables

$$p(y) = \phi^y(1-\phi)^{1-y}$$

$$p(x|y=0) = \frac{1}{(2\pi)^{d/2}}\frac{1}{|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu_0)^\top\Sigma^{-1}(x-\mu_0)\right)$$

$$p(x|y=1) = \frac{1}{(2\pi)^{d/2}}\frac{1}{|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu_1)^\top\Sigma^{-1}(x-\mu_1)\right)$$

MLE

- Given data D, we want to maximize likelihood!

$$\operatorname{argmax} \log p(D|\phi, \mu_0, \mu_1, \Sigma) =$$

MLE

- Given data D, we want to maximize likelihood!

$$\operatorname{argmax} \log p(D; \phi, \mu_0, \mu_1, \Sigma,) = \log \prod_{i=1}^{N} p(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma,)$$

$$= \log \prod_{i=1}^{N} p\left(x^{(i)} | y^{(i)}; \mu_0, \mu_1, \Sigma\right) p(y^{(i)}; \phi)$$

$$= \sum_{i=1}^{N} \log p\left(x^{(i)} | y^{(i)}; \mu_0, \mu_1, \Sigma\right) + \sum_{i=1}^{N} \log p(y^{(i)}; \phi)$$

MLE

- Given data D, we want to maximize likelihood!

$$\frac{\partial \log p}{\partial \phi} =$$

MLE

- Given data D, we want to maximize likelihood!

$$\frac{\partial \log p}{\partial \phi} = \frac{\partial}{\partial \phi}\left(\sum_{i=1}^{N} y^{(i)} \log \phi + \left(1 - y^{(i)}\right) \log(1-\phi)\right)$$

$$= \frac{1}{\phi}\sum_{i=1}^{N} y^{(i)} - \frac{1}{1-\phi}\sum_{i=1}^{N}\left(1 - y^{(i)}\right) = 0$$

$$\frac{1}{\phi}\sum_{i=1}^{N} y^{(i)} = \frac{1}{1-\phi}\sum_{i=1}^{N}\left(1 - y^{(i)}\right) \qquad \sum_{i=1}^{N} y^{(i)} = N\phi \qquad \phi^* = \frac{1}{N}\sum_{i=1}^{N} y^{(i)}$$

MLE

- Given data D, we want to maximize likelihood!

$$\frac{\partial \log p}{\partial \mu_1} =$$

MLE

- Given data D, we want to maximize likelihood!

$$\frac{\partial \log p}{\partial \mu_1} = \frac{\partial}{\partial \mu_1} \sum_{i=1}^{N} \log p\left(x^{(i)} | y^{(i)}; \mu_0, \mu_1, \Sigma\right)$$

$$= \frac{\partial}{\partial \mu_1} \sum_{i \in \{j | y^{(j)} = 1\}} \log p\left(x^{(i)} | y^{(i)}; \mu_1, \Sigma\right)$$

$$\mu_1^* = \frac{\sum_{i=1}^{N} 1\{y^{(i)} = 1\}\, x^{(i)}}{\sum_{i=1}^{N} 1\{y^{(i)} = 1\}} \qquad \mu_0^* = \frac{\sum_{i=1}^{N} 1\{y^{(i)} = 0\}\, x^{(i)}}{\sum_{i=1}^{N} 1\{y^{(i)} = 0\}}$$

MLE

- Given data D, we want to maximize likelihood!

$$\frac{\partial \log p}{\partial \Sigma} =$$

MLE

- Given data D, we want to maximize likelihood!

$$\frac{\partial \log p}{\partial \Sigma} = \frac{\partial}{\partial \Sigma} \sum_{i=1}^{N} \log p\left(x^{(i)} | y^{(i)}; \mu_0, \mu_1, \Sigma\right)$$

$$\Sigma^* = \frac{1}{N} \sum_{i=1}^{N} \left(x^{(i)} - \mu_{y^{(i)}}\right)\left(x^{(i)} - \mu_{y^{(i)}}\right)^{\top}$$

Testing

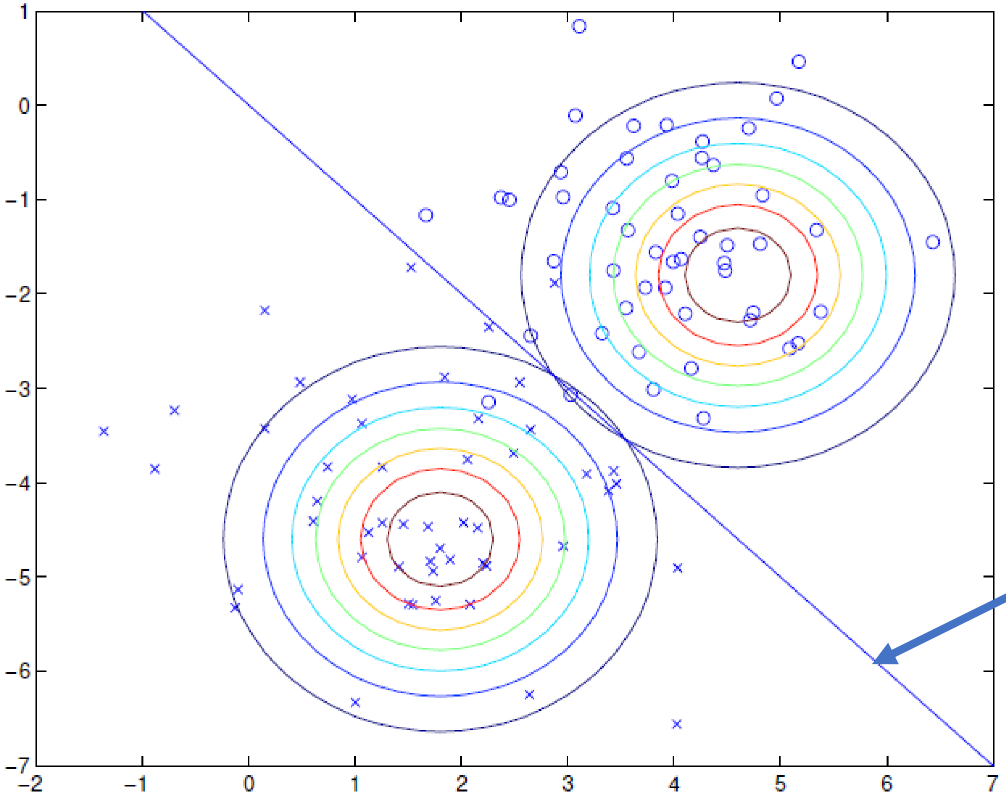
- Given a new data x^{new}

$$\underset{y}{\operatorname{argmax}}\, p(y|x^{new}) = \underset{y}{\operatorname{argmax}} \frac{p(x^{new}|y)p(y)}{p(x)} = \underset{y}{\operatorname{argmax}}\, p(x^{new}|y)p(y)$$

Compute $\quad p(x^{new}|y=0), p(y=0), p(x^{new}|y=1), p(y=1)$

Linear Decision Boundary

- Linear Discriminant Analysis



$$p(y=1|x)=0.5$$
$$=p(y=0|x)$$

Linear Decision Boundary

- Linear Discriminant Analysis

$$\log p(y = 1|x) = \log p(y = 0|x)$$

Linear Decision Boundary

- Linear Discriminant Analysis

$$\log p(y = 1|x) = \log p(y = 0|x)$$

$$(x - \mu_0)^\top \Sigma^{-1} (x - \mu_0) = (x - \mu_1)^\top \Sigma^{-1} (x - \mu_1) + \mathrm{const}$$

$$x^\top \Sigma^{-1} x - 2\mu_0^\top \Sigma^{-1} x + \mu_0^\top \mu_0 = x^\top \Sigma^{-1} x - 2\mu_1^\top \Sigma^{-1} x + \mu_1^\top \mu_1 + \mathrm{const}$$

$$-2(\mu_0^\top \Sigma^{-1} + \mu_1^\top \Sigma^{-1})x + \mu_0^\top \mu_0 - \mu_1^\top \mu_1 + \mathrm{const} = 0$$

Quadratic Decision Boundary

- Quadratic Discriminant Analysis, no shared covariance!

$$\log p(y = 1|x) = \log p(y = 0|x)$$

Quadratic Decision Boundary

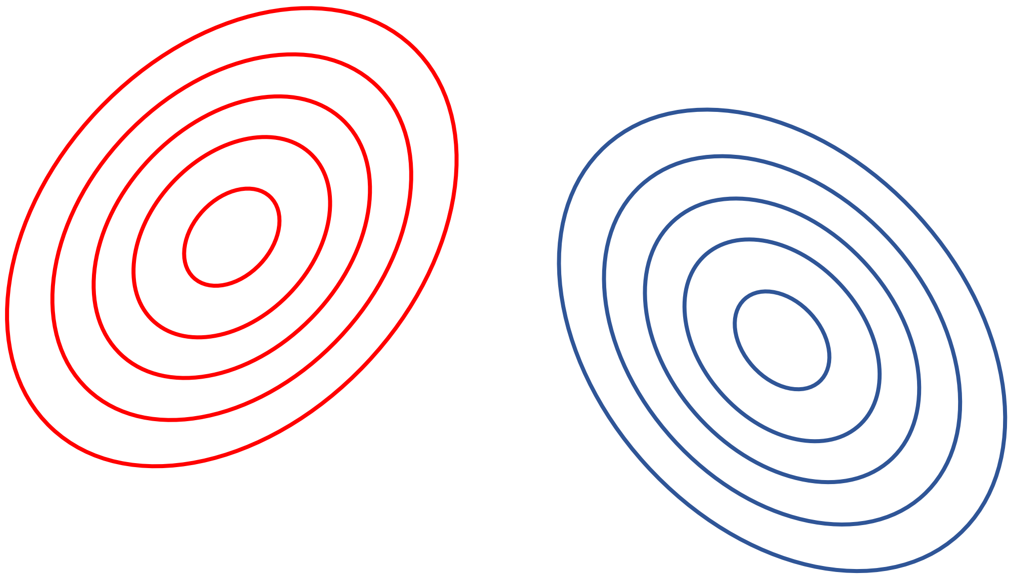
- Quadratic Discriminant Analysis, no shared covariance!

$$\log p(y = 1|x) = \log p(y = 0|x)$$

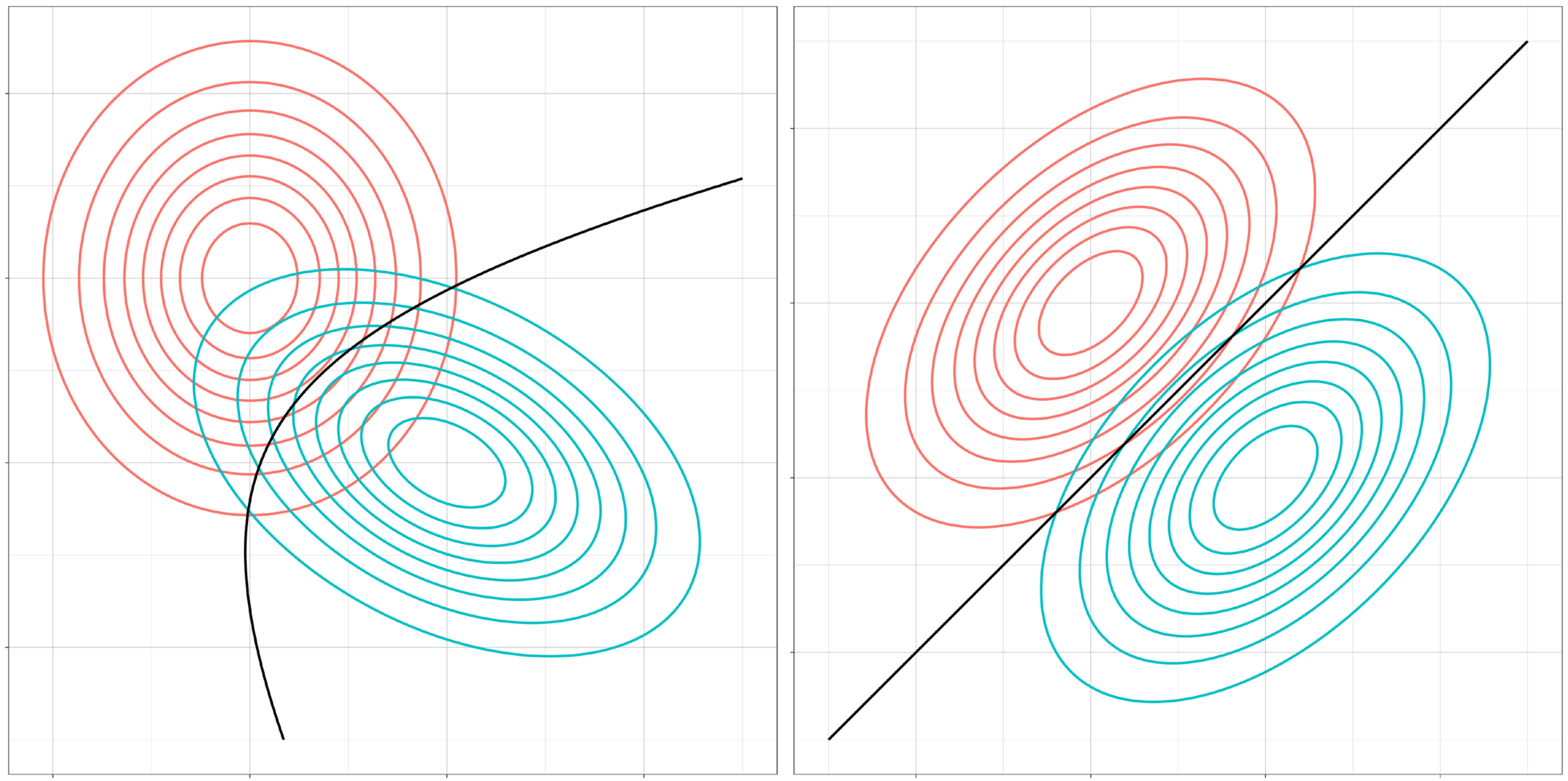
$$\log \frac{1}{|\Sigma_0|^{\frac{1}{2}}} + (x - \mu_0)^\top \Sigma_0^{-1} (x - \mu_0) = \log \frac{1}{|\Sigma_1|^{\frac{1}{2}}} + (x - \mu_1)^\top \Sigma_1^{-1} (x - \mu_1) + \mathrm{const}$$

Quadratic Decision Boundary

- Quadratic Discriminant Analysis, no shared covariance!

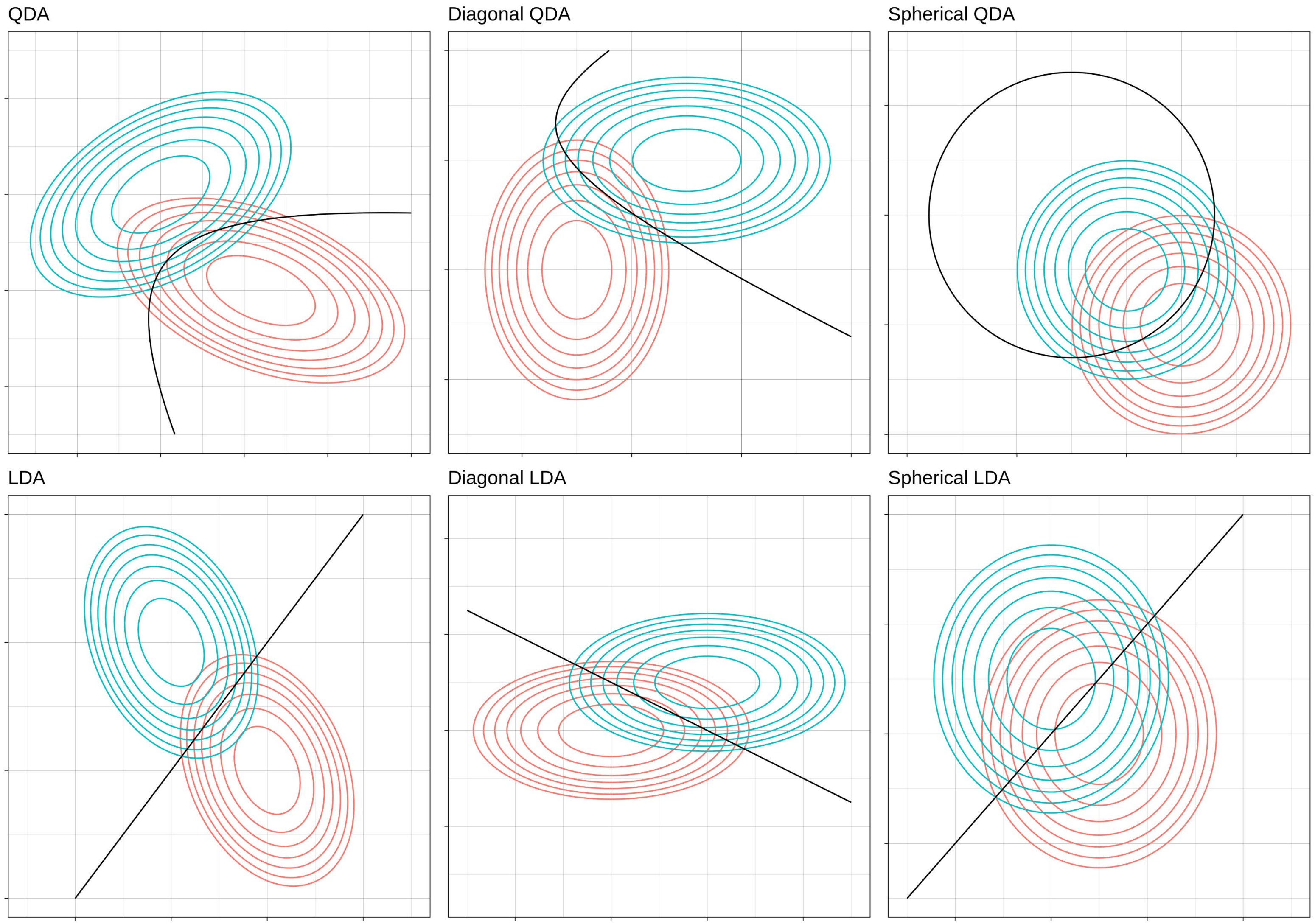


LDA vs QDA



[Six Varieties of Gaussian Discriminant Analysis (mathformachines.com)]

LDA vs QDA



[Six Varieties of Gaussian Discriminant Analysis (mathformachines.com)](https://mathformachines.com)

GDA vs Logistic Regression

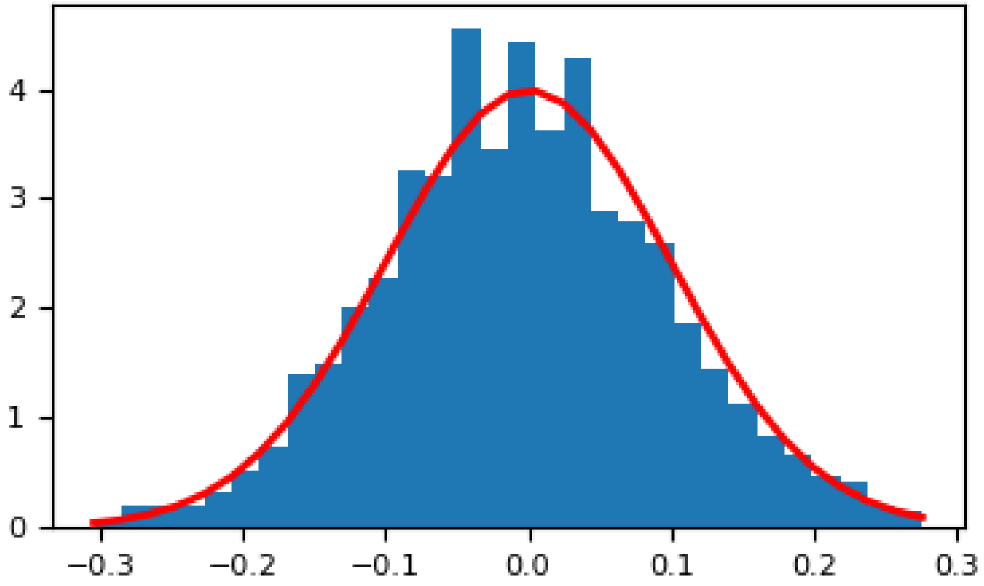
- GDA makes stronger assumption: $p(x|y)$ is a gaussian
- If the assumption is true, then GDA is "asymptotically efficient"
  - The best possible model when $\mathrm{N} \to \infty$
- When data is not Gaussian, logistic regression beats GDA when N is large
- GDA is usually better than logistic regression when N is small
- GDA is a generative model, so we can sample!

How To Generate Samples?

How to Generate Samples?

numpy.random.normal

```
random.normal(loc=0.0, scale=1.0, size=None)
```



numpy.random.normal — NumPy v1.26 Manual

Random Number Generators

- Sampling from a uniform distribution over [0,1)

numpy.random.rand

```
random.rand(d0, d1, ..., dn)
```

Pseudo Random Number Generators

- Linear Congruential Generators

$$X_{n+1} = (aX_n + c) \bmod d$$

$$\text{seed} = X_0 = 1$$
$$a = 5, c = 3, d = 9$$

$$0 < d$$
$$0 < a < d$$
$$0 \leq c < d$$

$$X_0 = 1$$

$$X_1 = (5 \cdot 1 + 3) \bmod 9 = 8$$

$$X_2 = (5 \cdot 8 + 3) \bmod 9 = 7$$

$$X_3 = (5 \cdot 7 + 3) \bmod 9 = 2$$

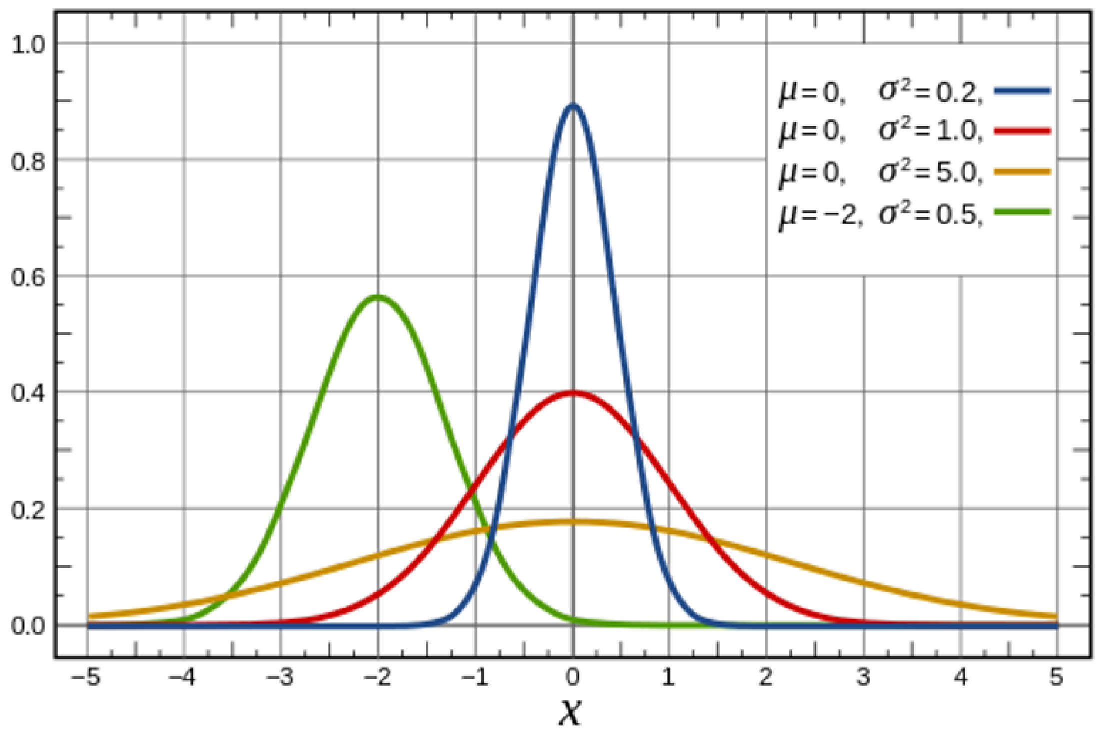
$$X_4 = (5 \cdot 2 + 3) \bmod 9 = 4$$

…

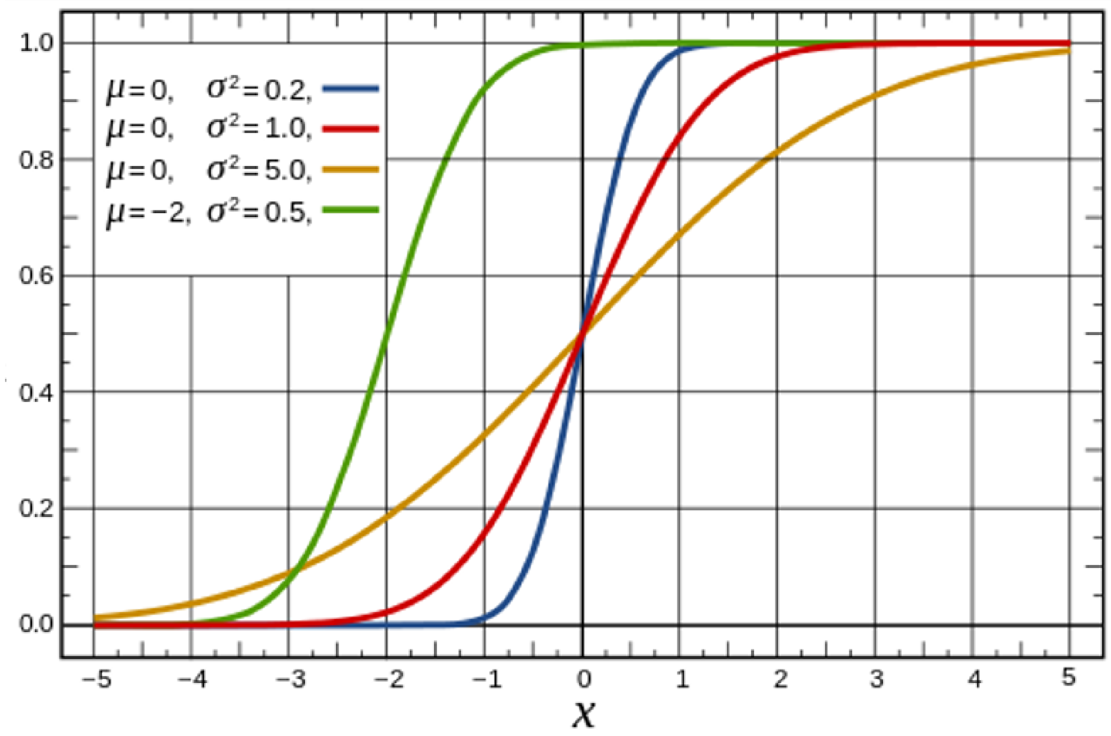
Sampling From Gaussian Distribution

- Inverse transform sampling

Probability Density Function



Cumulative Density Function



Sampling From Gaussian Distribution

- Standard Normal Distribution N(0, 1)

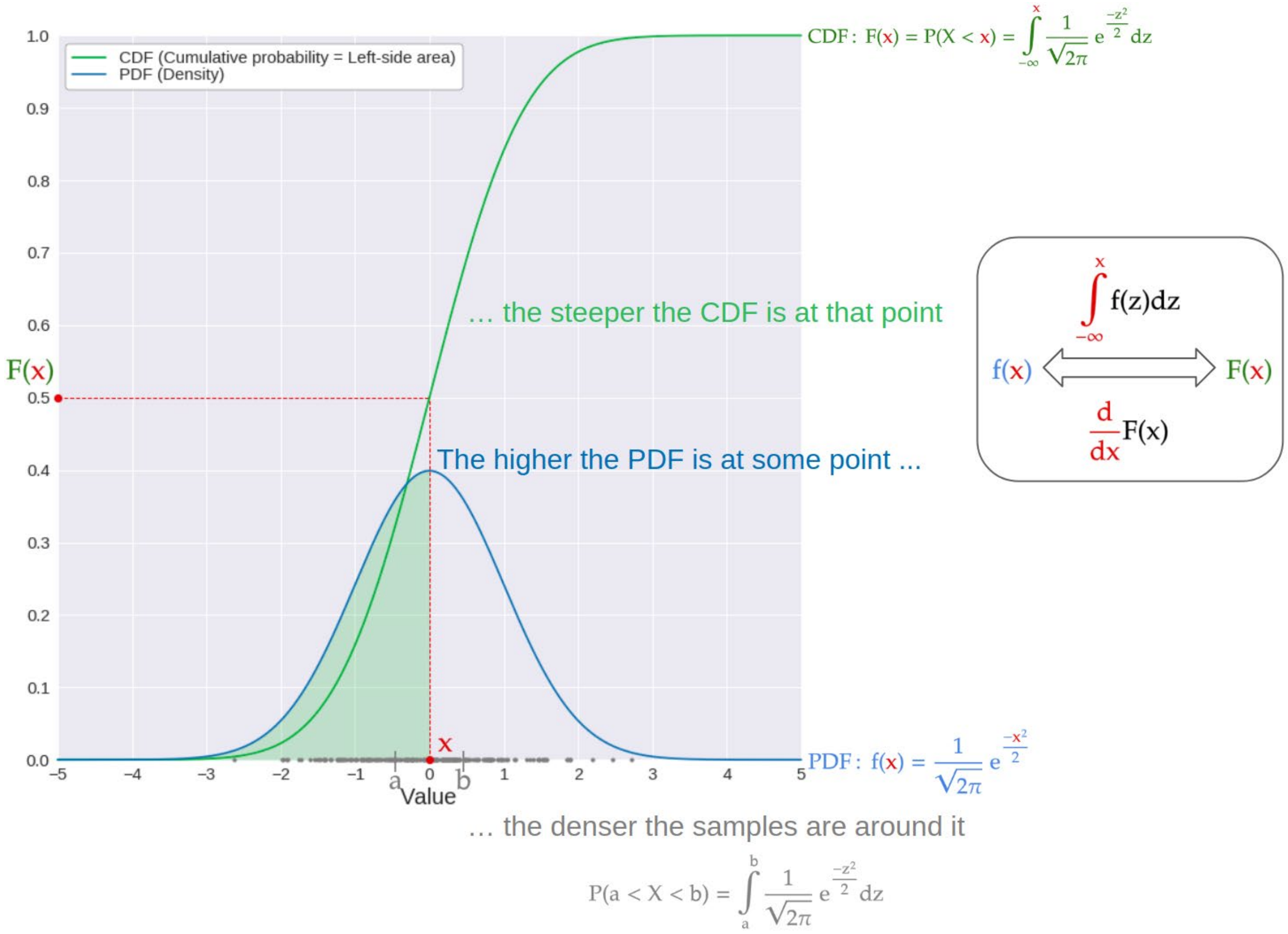
Probability Density Function

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Cumulative Density Function

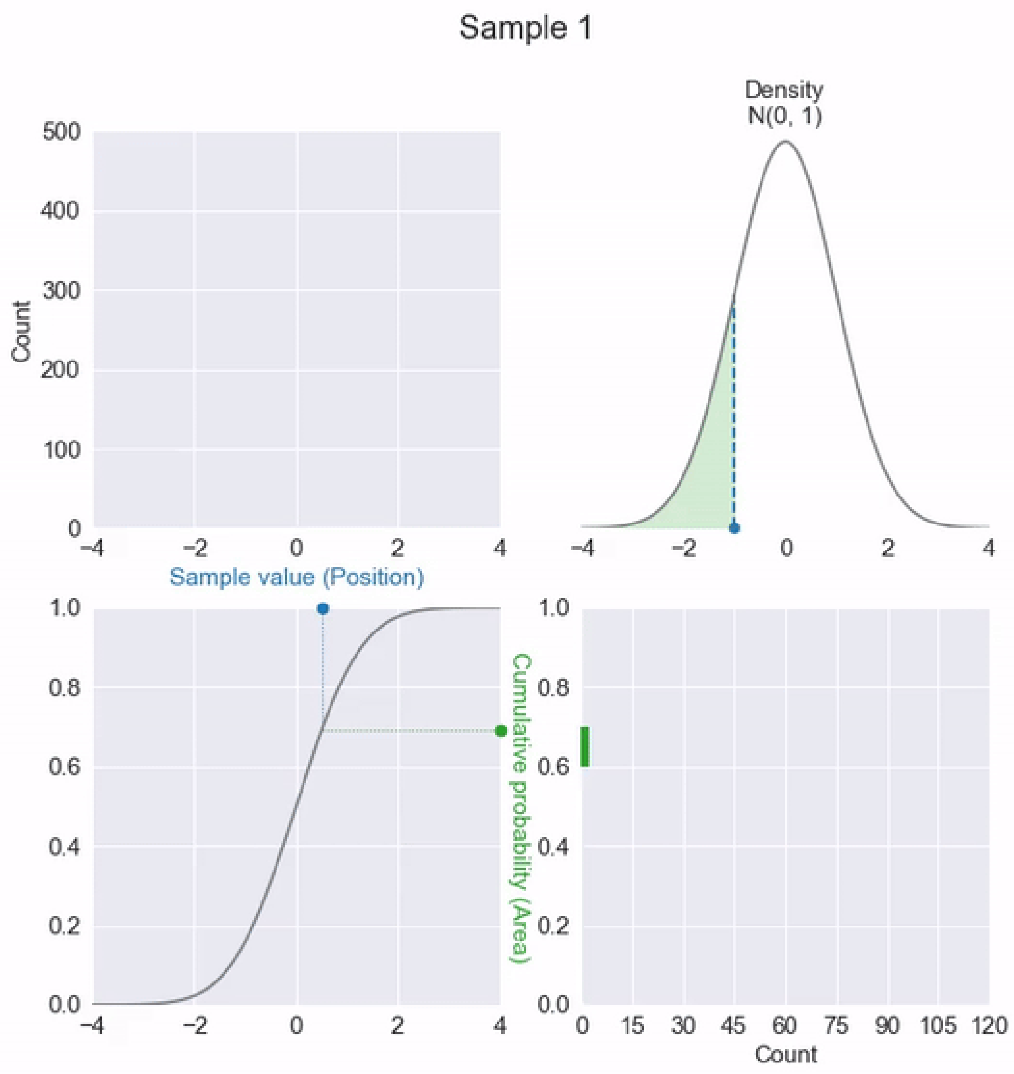
$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}} du$$

Inverse transform sampling



[How to generate Gaussian samples. Part 1: Inverse transform sampling | by Khanh Nguyen | MTI Technology | Medium]

Inverse transform sampling



Naïve Bayes

(Discrete Input Features)

Example: Spam Classification

$$\operatorname*{argmax}_y p(y|x) = \operatorname*{argmax}_y \frac{p(x|y)p(y)}{p(x)}$$

$$= \operatorname*{argmax}_y p(x|y)p(y)$$

Example: Spam Classification

- Each vocabulary is one feature dimension
- We encode each email as a feature vector $x \in \{0,1\}^{|V|}$
  - One-hot encoding
- $x_j = 1$, iff the vocabulary x_j appears in the email

$$x = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ \dots \\ 0 \end{bmatrix} \in \{0,1\}^{|V|}$$

a
Skku
University
Buy
He
She
…

$$\underset{y}{\operatorname{argmax}}\, p(x|y)p(y)$$

y: spam or not

x: input

09_naive_bayes.pdf (toronto.edu)

Example: Spam Classification

- We want to model the probability of any word x_j appearing in an email given the email is spam or not

09_naive_bayes.pdf (toronto.edu)

Issues

- What if |V| (the number of vocabulary) is large?
- Example: |V| =3
- 2^3 possible outcome, 2^3 classification

$$p(x|y)$$

(categorical distribution)

$$\begin{bmatrix}0\\0\\0\end{bmatrix} \quad \begin{bmatrix}0\\0\\1\end{bmatrix} \quad \begin{bmatrix}0\\1\\0\end{bmatrix} \quad \begin{bmatrix}0\\1\\1\end{bmatrix} \quad \begin{bmatrix}1\\0\\0\end{bmatrix} \quad \begin{bmatrix}1\\0\\1\end{bmatrix} \quad \begin{bmatrix}1\\1\\0\end{bmatrix} \quad \begin{bmatrix}1\\1\\1\end{bmatrix}$$

[09_naive_bayes.pdf (toronto.edu)](09_naive_bayes.pdf)

Example: Spam Classification

- Naïve Bayes
- $p(x_j|y = k)$ is a Bernoulli distribution

$$p(x|y = k) = \prod_{j=1}^{|V|} p(x_j|y = k)$$

- Not a right assumption in practice
  - If y=1 (spam), then, knowledge of 'buy' presence does not affect on your beliefs about other words, e.g., 'price'

09_naive_bayes.pdf (toronto.edu)

Example: Spam Classification

- Both $p(x_j|y=1)$ and $p(y=1)$ are Bernoulli distributions

$$p(x|y=k) = \prod_{j=1}^{|V|} p(x_j|y=k) \qquad \phi \in [0,1]^{|V|\times 2} \qquad \theta \in [0,1]$$

$$p(x_j=1|y=1) = \phi_{j1}$$

$$p(x_j=0|y=1) = 1-\phi_{j1}$$

$$p(x_j=1|y=0) = \phi_{j0}$$

$$p(x_j=0|y=0) = 1-\phi_{j0}$$

$$p(y=1) = \theta$$

$$p(y=0) = 1-\theta$$

MLE

$$\log L(\phi, \theta) = \log \prod_{i=1}^{N} p\left(x^{(i)}, y^{(i)}; \phi, \theta\right)$$

MLE

$$\begin{aligned}
\log L(\phi, \theta) &= \log \prod_{i=1}^{N} p\left(x^{(i)}, y^{(i)}; \phi, \theta\right) \\
&= \log \prod_{i=1}^{N} p\left(x^{(i)} | y^{(i)}; \phi\right) p\left(y^{(i)}; \theta\right) \\
&= \log \prod_{i=1}^{N} p\left(y^{(i)}; \theta\right) \prod_{j=1}^{|V|} p\left(x_j^{(i)} | y^{(i)}; \phi\right) \\
&= \sum_{i=1}^{N} \log p\left(y^{(i)}; \theta\right) + \sum_{i=1}^{N} \sum_{j=1}^{|V|} \log p\left(x_j^{(i)} | y^{(i)}; \phi\right)
\end{aligned}$$

MLE

$$\frac{\partial L}{\partial \phi_{l1}} = \frac{\partial}{\partial \phi_{l1}} \sum_{i=1}^{N} \log p\left(y^{(i)}; \theta\right) + \sum_{i=1}^{N} \sum_{j=1}^{|V|} \log p\left(x_j^{(i)} | y^{(i)}; \phi\right)$$

$$= \frac{\partial}{\partial \phi_{l1}} \sum_{i \in \{k | y^{(k)}=1\}} \sum_{j=1}^{|V|} \log p\left(x_j^{(i)} | y^{(i)}; \phi\right)$$

$$= \frac{\partial}{\partial \phi_{l1}} \sum_{i \in \{k | y^{(k)}=1\}} \sum_{j=1}^{|V|} \log\left(\phi_{j1}^{x_j^{(i)}} (1 - \phi_{j1})^{\left(1 - x_j^{(i)}\right)}\right) = \frac{\partial}{\partial \phi_{l1}} \sum_{i \in \{k | y^{(k)}=1\}} \log\left(\phi_{l1}^{x_l^{(i)}} (1 - \phi_{l1})^{\left(1 - x_l^{(i)}\right)}\right)$$

$$= \frac{\partial}{\partial \phi_{l1}} \sum_{i \in \{k | y^{(k)}=1\}} x_l^{(i)} \log \phi_{l1} + \left(1 - x_l^{(i)}\right) \log(1 - \phi_{l1})$$

$$= \sum_{i \in \{k | y^{(k)}=1\}} \frac{x_l^{(i)}}{\phi_{l1}} - \frac{1 - x_l^{(i)}}{(1 - \phi_{l1})} = 0$$

MLE

$$\sum_{i\in\{k|y^{(k)}=1\}} \frac{x_l^{(i)}}{\phi_{l1}} = \sum_{i\in\{k|y^{(k)}=1\}} \frac{1-x_l^{(i)}}{(1-\phi_{l1})}$$

$$\frac{1}{\phi_{l1}} \sum_{i\in\{k|y^{(k)}=1\}} x_l^{(i)} = \frac{1}{(1-\phi_{l1})} \sum_{i\in\{k|y^{(k)}=1\}} 1-x_l^{(i)}$$

$$(1-\phi_{l1}) \sum_{i\in\{k|y^{(k)}=1\}} x_l^{(i)} = \phi_{l1} \sum_{i\in\{k|y^{(k)}=1\}} 1-x_l^{(i)}$$

$$\sum_{i\in\{k|y^{(k)}=1\}} x_l^{(i)} = \phi_{l1} \sum_{i\in\{k|y^{(k)}=1\}} 1$$

$$\sum_{i\in\{k|y^{(k)}=1\}} x_l^{(i)} = \phi_{l1} \sum_{i=1}^{N} 1\{y^{(i)} = 1\}$$

$$\phi_{l1}^* = \frac{\sum_{i=1}^{N} 1\{y^{(i)} = 1\} x_l^{(i)}}{\sum_{i=1}^{N} 1\{y^{(i)} = 1\}} = \frac{\sum_{i=1}^{N} 1\{y^{(i)} = 1\} x_l^{(i)}}{N_1}$$

MLE

$$\sum_{i\in\{k|y^{(k)}=1\}} \frac{x_l^{(i)}}{\phi_{l1}} = \sum_{i\in\{k|y^{(k)}=1\}} \frac{1-x_l^{(i)}}{(1-\phi_{l1})}$$

$$\frac{1}{\phi_{l1}} \sum_{i\in\{k|y^{(k)}=1\}} x_l^{(i)} = \frac{1}{(1-\phi_{l1})} \sum_{i\in\{k|y^{(k)}=1\}} 1-x_l^{(i)}$$

$$(1-\phi_{l1}) \sum_{i\in\{k|y^{(k)}=1\}} x_l^{(i)} = \phi_{l1} \sum_{i\in\{k|y^{(k)}=1\}} 1-x_l^{(i)}$$

$$\sum_{i\in\{k|y^{(k)}=1\}} x_l^{(i)} = \phi_{l1} \sum_{i\in\{k|y^{(k)}=1\}} 1$$

$$\sum_{i\in\{k|y^{(k)}=1\}} x_l^{(i)} = \phi_{l1} \sum_{i=1}^{N} 1\{y^{(i)} = 1\}$$

$$\phi_{l1}^* = \frac{\sum_{i=1}^N 1\{y^{(i)} = 1\}x_l^{(i)}}{\sum_{i=1}^N 1\{y^{(i)} = 1\}} = \frac{\sum_{i=1}^N 1\{y^{(i)} = 1\}x_l^{(i)}}{N_1}$$

$$\phi_{l0}^* = \frac{\sum_{i=1}^N 1\{y^{(i)} = 0\}x_l^{(i)}}{\sum_{i=1}^N 1\{y^{(i)} = 0\}} = \frac{\sum_{i=1}^N 1\{y^{(i)} = 0\}x_l^{(i)}}{N_0}$$

MLE

$$\sum_{i\in\{k|y^{(k)}=1\}} \frac{x_l^{(i)}}{\phi_{l1}} = \sum_{i\in\{k|y^{(k)}=1\}} \frac{1-x_l^{(i)}}{(1-\phi_{l1})}$$

$$\frac{1}{\phi_{l1}} \sum_{i\in\{k|y^{(k)}=1\}} x_l^{(i)} = \frac{1}{(1-\phi_{l1})} \sum_{i\in\{k|y^{(k)}=1\}} 1-x_l^{(i)}$$

$$(1-\phi_{l1}) \sum_{i\in\{k|y^{(k)}=1\}} x_l^{(i)} = \phi_{l1} \sum_{i\in\{k|y^{(k)}=1\}} 1-x_l^{(i)}$$

$$\sum_{i\in\{k|y^{(k)}=1\}} x_l^{(i)} = \phi_{l1} \sum_{i\in\{k|y^{(k)}=1\}} 1$$

$$\sum_{i\in\{k|y^{(k)}=1\}} x_l^{(i)} = \phi_{l1} \sum_{i=1}^{N} 1\{y^{(i)} = 1\}$$

$$\phi_{l1}^* = \frac{\sum_{i=1}^N 1\{y^{(i)} = 1\} x_l^{(i)}}{\sum_{i=1}^N 1\{y^{(i)} = 1\}} = \frac{\sum_{i=1}^N 1\{y^{(i)} = 1\} x_l^{(i)}}{N_1}$$

$$\phi_{l0}^* = \frac{\sum_{i=1}^N 1\{y^{(i)} = 0\} x_l^{(i)}}{\sum_{i=1}^N 1\{y^{(i)} = 0\}} = \frac{\sum_{i=1}^N 1\{y^{(i)} = 0\} x_l^{(i)}}{N_0}$$

$$\phi_{lk}^* = \frac{\sum_{i=1}^N 1\{y^{(i)} = k\} x_l^{(i)}}{\sum_{i=1}^N 1\{y^{(i)} = k\}} = \frac{\sum_{i=1}^N 1\{y^{(i)} = k\} x_l^{(i)}}{N_k}$$

MLE

$$\frac{\partial L}{\partial \theta} \log L(\phi, \theta) = \frac{\partial L}{\partial \theta} \sum_{i=1}^{N} \log p\left(y^{(i)}; \theta\right) + \sum_{i=1}^{N} \sum_{j=1}^{|V|} \log p\left(x_j^{(i)} | y^{(i)}; \phi\right)$$

$$= \frac{\partial L}{\partial \theta} \sum_{i=1}^{N} \log p\left(y^{(i)}; \theta\right)$$

$$\theta^* = \frac{\sum_{i=1}^{N} 1\{y^{(i)} = 1\}}{N} = \frac{N_1}{N} \qquad \theta_k^* = \frac{\sum_{i=1}^{N} 1\{y^{(i)} = k\}}{N} = \frac{N_k}{N}$$

Testing

$$\underset{k}{\operatorname{argmax}}\, p(y = k) \prod_{j=1}^{d} p\left(x_j^{new} | y = k\right)$$

Testing

$$\underset{k}{\operatorname{argmax}}\, p(y=k) \prod_{j=1}^{d} p\left(x_j^{new} | y=k\right)$$

$$\underset{k}{\operatorname{argmax}}\, \theta_k^* \prod_{j=1}^{d} {\phi_{jk}^*}^{x_j^{new}} \left(1-\phi_{jk}^*\right)^{1-x_j^{new}}$$

Laplace Smoothing

- What if we have not seen a word "skku" before?
- Then,

$$p(x_{30}|y=1;\phi)=\phi_{30,1}=0$$

$$p(y=1|x)=\frac{p(y=1)\prod_{j=1}^{d}p\left(x_j|y=1\right)}{p(x)}$$

$$=\frac{p(y=1)\prod_{j=1}^{d}p\left(x_j|y=1\right)}{p(y=0)\prod_{j=1}^{d}p\left(x_j|y=0\right)+p(y=1)\prod_{j=1}^{d}p\left(x_j|y=1\right)}=\frac{0}{0}$$

Laplace Smoothing

- Statistically, it is a bad idea to say probability is ‘zero’ just because you haven’t seen it!
- So, add ‘1’ to numerator, K to denominator

$$\theta_k^* = \frac{\sum_{i=1}^{N} 1\{y^{(i)} = k\}}{N} \longrightarrow \theta_k^* = \frac{1 + \sum_{i=1}^{N} 1\{y^{(i)} = k\}}{K + N}$$

$$\phi_{lk}^* = \frac{\sum_{i=1}^{N} 1\{y^{(i)} = 1\} x_l^{(i)}}{N_k} \longrightarrow \phi_{lk}^* = \frac{1 + \sum_{i=1}^{N} 1\{y^{(i)} = 1\} x_l^{(i)}}{K + N_k}$$