CS229 Note Lecture 05

Generative Learning Algorithms

Discriminative

- Learns $P(y \mid x)$
- Or learns $h_\theta(x) \in \{0, 1\}$ directly.

Generative

$$P(x \mid y) \qquad P(y)$$

A generative model builds a probabilistic model for what the feature looks like conditioned on the class label.

By **Bayes Rule**, obviously,

$$P(y = 1 \mid x) = \frac{P(x \mid y = 1)P(x)}{P(x)}$$

and

$$P(x) = P(y = 0 \mid x)P(x) + P(y = 1 \mid x)P(x)$$

Gaussian Discriminant Analysis (GDA)

Multivariate Gaussian Distribution

Assume $x \in \mathbb{R}^n$, continued-values.

Another assumption is

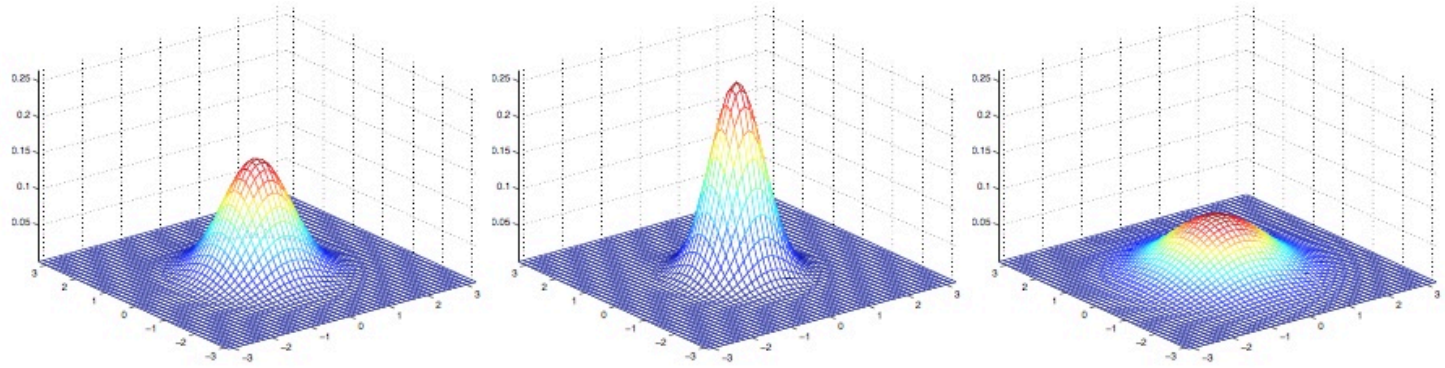
$$P(x \mid y) \sim \mathcal{N}(\mu, \sigma^2)$$

If $z \in \mathcal{N}(\vec{\mu}, \Sigma)$, so the density of z is

$$\frac{1}{\sqrt{(2\pi)^n|\Sigma|}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

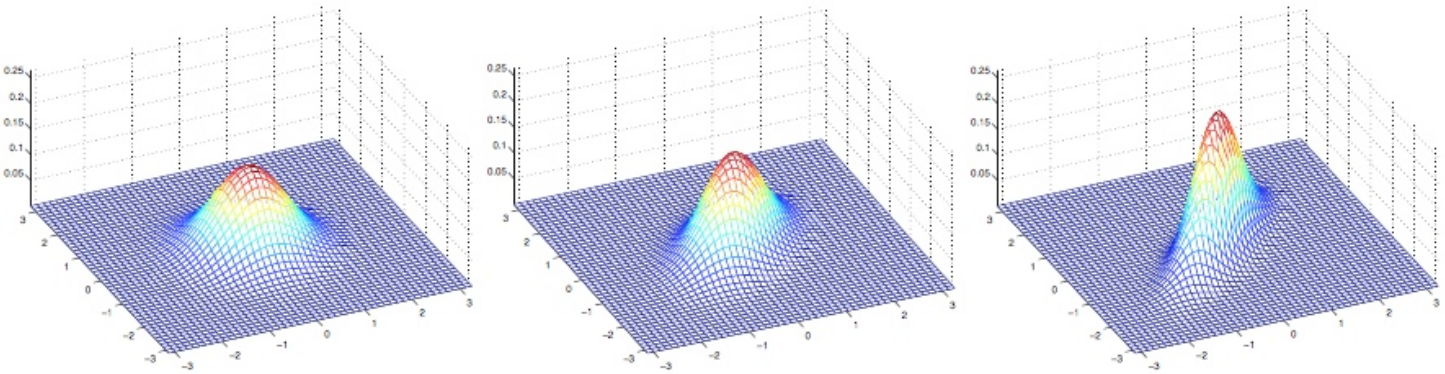
where μ is the mean of the Gaussian, and matrix \Siama is the **Covariance Matrix** and so Σ will be equal to $E[(x-\mu)(x-\mu)^T]$.

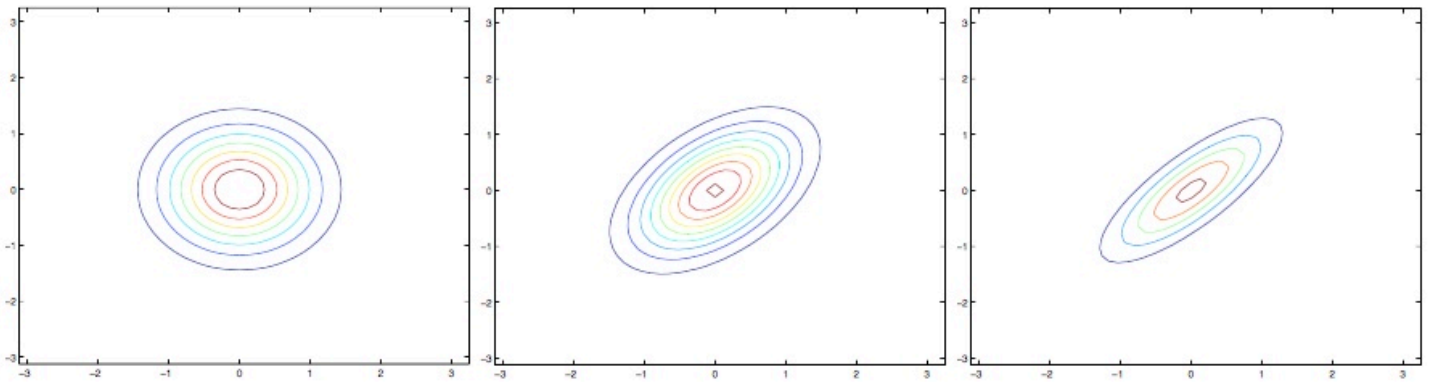
$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$



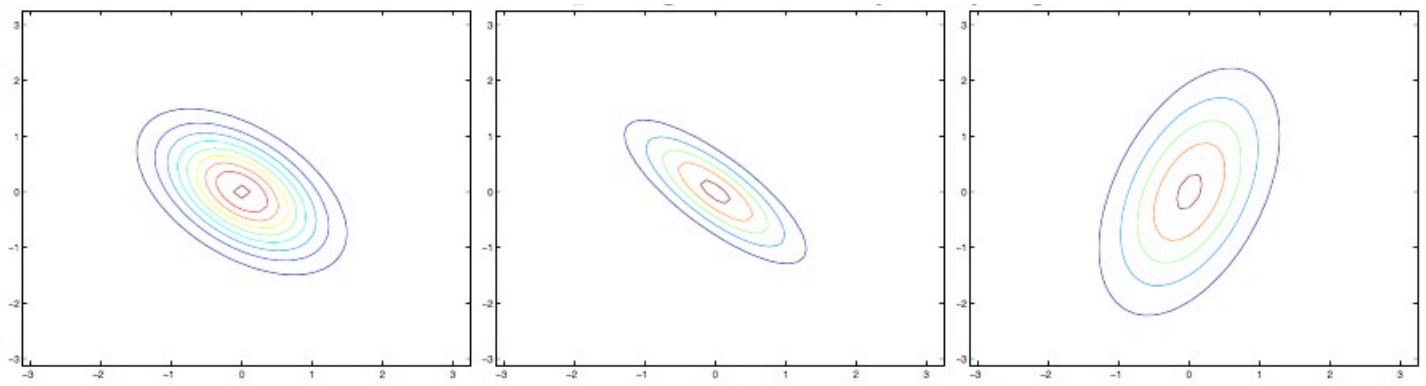
)

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

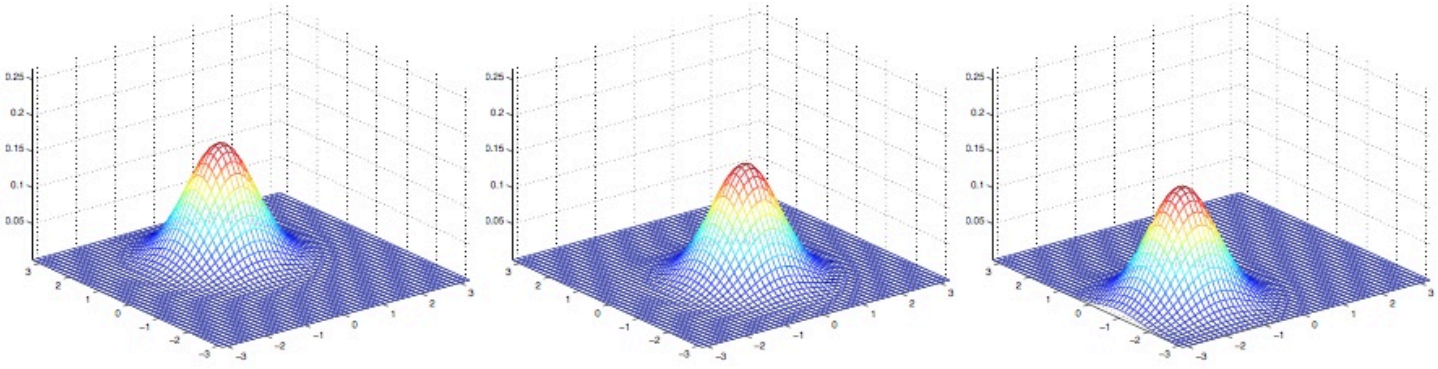


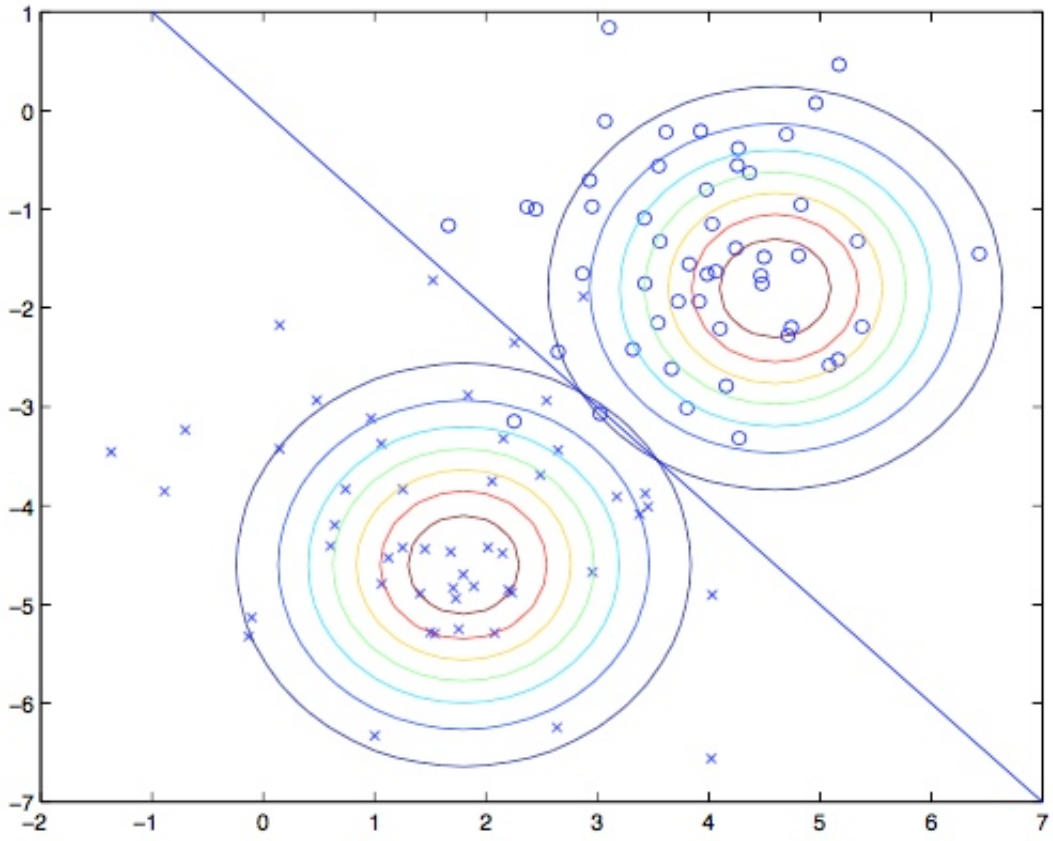


$$\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 3 & 0.8 \\ 0.8 & 3 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \mu = \begin{bmatrix} -0.5 \\ 0 \end{bmatrix} \qquad \mu = \begin{bmatrix} -1 \\ -1.5 \end{bmatrix}$$





What we are going to do is, just looking at only the positive examples, we are gonna fit a Gaussian distribution to the positive examples. Then we will look at the negative examples, and fit a Gaussian distribution. And together these Gaussian densities define a separator for

these two classes.

Gaussian Discriminant Analysis Model

Put $p(y)$ into **Bernoulli Distribution**

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

Model $P(x)$ given $Y = 0$ as a Gaussian

$$P(x \mid y=0) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(x-\mu_0)^T \Sigma^{-1}(x-\mu_0)}$$

$$P(x \mid y=1) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)}$$

The parameters are

$$\phi, \mu_0, \mu_1, \Sigma$$

so the log likelihood of the parameters is

$$\begin{aligned} \ell(\phi, \mu_0, \mu_1, \Sigma) &= log \prod_{i=1}^{m} P(x^{(i)}, y^{(i)}) \\ &= log \prod_{i=1}^{m} P(x^{(i)} \mid y^{(i)}) \cdot P(y^{(i)}) \end{aligned}$$

The equation above often called **Joint Likelihood**.

So given the train sets and using the GDA model, to fix the parameters of the model, we will do likelihood estimation as usual, maximize l with $\phi, \mu_0, \mu_1, \Sigma$

$$\begin{aligned} \phi &= \frac{\sum_i y^{(i)}}{m} = \frac{\sum_{i=1}^m 1\{y^{(i)}=1\}}{m} \\ \mu_0 &= \frac{\sum_{i=1}^m 1\{y^{(i)}=0\} x^{(i)}}{\sum_{i=1}^m 1\{y^{(i)}=0\}} \\ \mu_1 &= \frac{\sum_{i=1}^m 1\{y^{(i)}=1\} x^{(i)}}{\sum_{i=1}^m 1\{y^{(i)}=1\}} \\ \Sigma &= \frac{1}{m} \sum_{i=1}^{m} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T \end{aligned}$$

For μ_0, $\sum_{i=1}^{m} 1\{y^{(i)} = 0\}x^{(i)}$ is the sum of $x^{(i)}$ for which $y^{(i)} = 0$. And $\sum_{i=1}^{m} 1\{y^{(i)} = 0\}$ is the number of examples with label 0.

Having fit parameters to the data, now we need to make prediction.

$$\begin{aligned}\arg\max_y P(y \mid x) &= \arg\max_y \frac{P(x \mid y)P(y)}{P(x)} \\ &= \arg\max_y P(x \mid y)P(y)\end{aligned}$$

If $P(y)$ is uniform, in other words if each of our constants is equally likely, so if $P(y)$ takes the same value for all values of y, then this is just

$$\arg\max_y P(x \mid y)$$

Generative vs. Discriminative Learning Algorithm

Assume $x \mid y \sim \mathcal{N}$, that gives logistic posterior for $P(y = 1 \mid x)$, and it turns out this implication in the opposite direction does not hold true.

Actually, if we assume

$$\begin{aligned}x \mid y = 1 &\sim Poisson(\lambda_1) \\ x \mid y = 0 &\sim Poisson(\lambda_0)\end{aligned}$$

then that also implies that $P(y \mid x)$ is logistic.

Gaussian discriminant analysis makes a stronger assumption that $x \mid y \sim \mathcal{N}$, so when this assumption is true or approximately holds, the algorithm will do better than logistic regression.Conversely, if not sure what $x|y$ is, then logistic regression, the discriminant algorithm may do better.

The real advantage of generative learning algorithm is often that it requires less data, and data is never really exactly
Gaussian because data is often approximately Gaussian. It turns out that generative learning algorithm often do surprisingly well even when those modeling assumption are not met. One other tradeoff is that by making stronger assumption about the data, Gaussian discriminant analysis often needs less data in order to fit an okay model.

In contrast, logistic regression by making less assumption is more robust to modeling assumption, but sometimes it takes a slightly larger training set to fit than Gaussian discriminant analysis.

General Version

Assume

$$x \mid y = 1 \sim ExpFamily(\eta_1)$$
$$x \mid y = 0 \sim ExpFamily(\eta_0)$$

this implies that $P(y = 1 \mid x)$ is also logistic.

Naive Bayes (Also Generative Learning Algorithm)

Represent a email using a feature vector X

$$X = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} a \\ aardvark \\ ausworth \\ \vdots \\ buy \\ \vdots \\ cs229 \\ \vdots \\ zymurgy \end{matrix}$$

Assume x_i's are **conditionally independent** given y. Condition independent means

$$\begin{aligned} P(x_1, x_2, \dots, x_n \mid y) &= P(x_1 \mid y)P(x_2 \mid y, x_1)P(x_3 \mid y, x_1, x_2) \cdots \\ &= P(x_1 \mid y)P(x_2 \mid y)P(x_3 \mid y) \cdots P(x_n \mid y) \\ &= \prod_{i=1}^{n} P(x_i \mid y) \end{aligned}$$

The parameters of the model are

$$\begin{aligned} \phi_{j|y=1} &= P(x_i = 1 \mid y = 1) \\ \phi_{j|y=0} &= P(x_i = 1 \mid y - 0) \\ \phi_y &= P(y = 1) \end{aligned}$$

Therefore, to fit the parameters of the model, the joint likelihood is

$$\mathcal{L}(\phi_y, \phi_{j|y=1}, \phi_{j|y=0}) = \prod_{i=1}^{m} P(x^{(i)}, y^{(i)})$$

Do likelihood estimate and we will find the maximum likelihood of the parameters are

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 1, y^{(i)} = 1\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}$$
$$\phi_y = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}{m}$$

By **Bayes Rule**

$$P(y \mid x) = P(x \mid y)P(y)$$

New Word

Given a new that never occured in previous email, if the word is 30000's

$$P(X_{30000} = 1 \mid y = 1) = 0$$
$$P(X_{30000} = 1 \mid y = 0) = 0$$

So when spam classifier goes to compute

$$P(y = 1 \mid x) = \frac{P(x \mid y = 1)P(y = 1)}{P(x \mid y = 1)P(y = 1) + P(x \mid y = 0)P(y = 0)}$$

the $P(x \mid y = 1) = \prod_{i=1}^{n} P(X_i \mid y = 1)$ because $P(X_{30000} = 1 \mid y = 1) = 0$, and in the same way it turns out the denominator will also be 0. So end up with

$$P(y = 1 \mid x) = \frac{0}{0 + 0} = \frac{0}{0}$$

and it doesn't make sense.

Laplace Smoothing

We can use **Laplace Smoothing** to fix this question

Date	*Team*	*Win*
2/8	Washington State	0
2/11	Washington	0
2/22	USC	0
2/24	UCLA	0
3/8	USC	0
3/15	Louisville	?

We estimate the probability of $P(y = 1)$, normally the maximum likelihood estimate is

$$P(y = 1) = \frac{\#\text{ " }1\text{ " }s}{\#\text{ " }0\text{ " }s + \#\text{ " }1\text{ " }s}$$

In the **Laplace Smoothing**, add 1 to all of these counts

$$\begin{aligned} P(y = 1) &= \frac{\#\text{ " }1\text{ " }s}{\#\text{ " }0\text{ " }s + \#\text{ " }1\text{ " }s} \\ &= \frac{(0 + 1)}{(5 + 1) + (0 + 1)} \\ &= \frac{1}{7} \end{aligned}$$

And more generally, if $y \in \{1, \ldots, k\}$

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

applies Laplace Smoothing on it

$$P(y = j) = \frac{\sum_{j=1}^{m} 1\{y^{(i)} = j\} + 1}{m + k}$$

So for Naive Bayes, what that gives us is

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{m} 1\{x^{(i)} = 1, y^{(i)} = 1\} + 1}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\} + 2}$$