CS 461: Machine Learning Principles

Class 9: Oct. 3

Fisher’s Discriminant Analysis and Logistic Regression

Instructor: Diana Kim

- Classification: Learning the discriminant functions $f_k(x)$

+ discriminant functions define the surfaces over the data / feature space.

$$C_k = \arg\max_k f_k(x)$$

: by using discriminant functions, we can assign a class to x

Example) three discriminant functions in the feature space of $\mathbb{R}^2$

1. $f_1(x_1, x_2) = x_2 - x_1 - 1$
2. $f_2(x_1, x_2) = x_2 + x_1 - 1$
3. $f_3(x_1, x_2) = x_2$

$$C_k = \arg\max_k f_k(x)$$

Q: Based on the discriminant functions above, assign a class for the points?

	f_1	f_2	f_3	class inference
$(-2,0)$	1	-3	0	1
$(0,0)$	-1	-1	0	3
$(2,0)$	-3	1	0	2

We are going learn
(1) how we learn **the discriminant functions** and
(2) how we define **the decision regions** based on discriminant functions.

For binary classification. (+ / −)

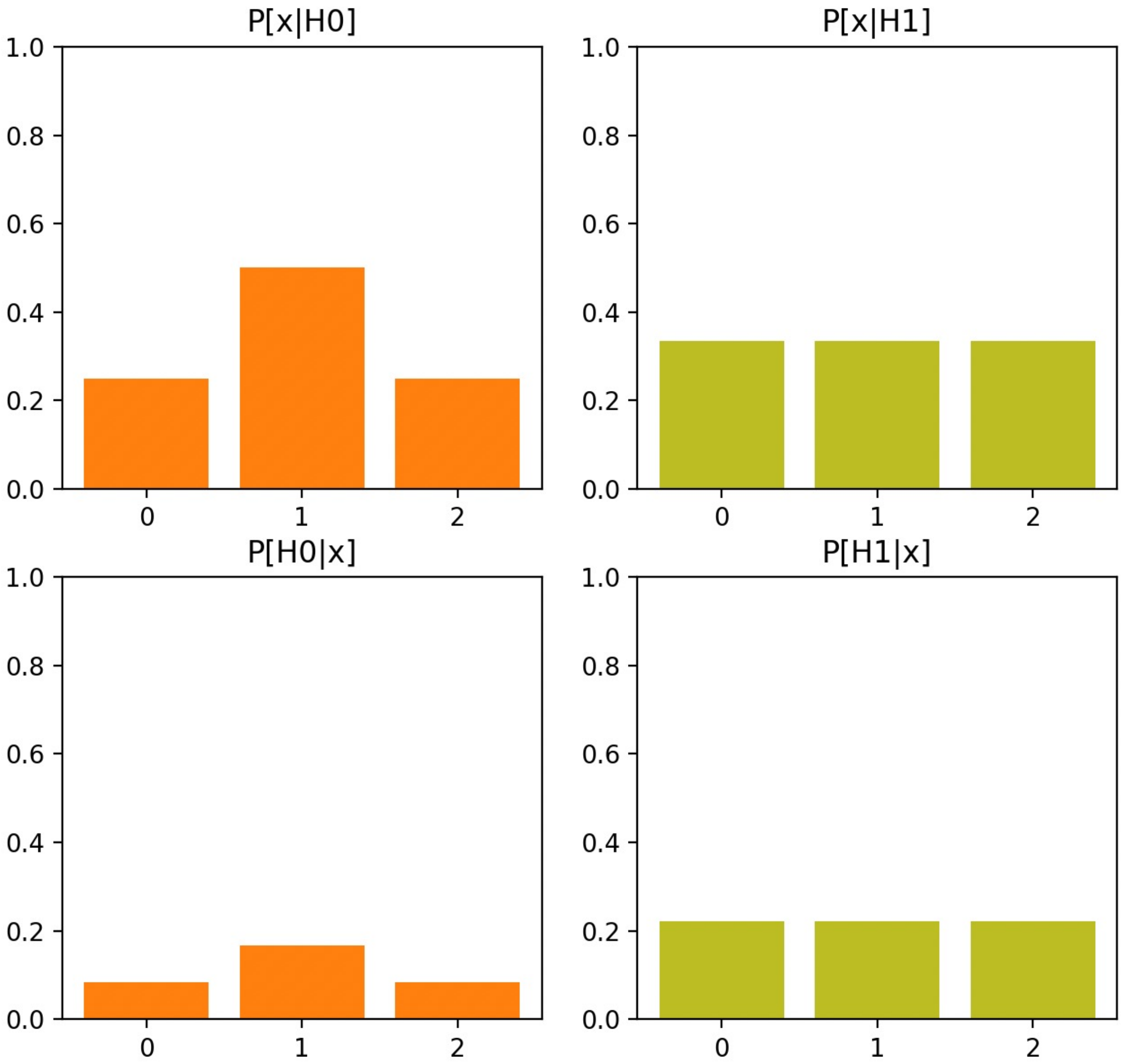
We need **two discriminant functions.**

MAP is one possible way to form the two discriminant functions.

$$\overbrace{\underbrace{p(x|H_0)}_{\text{likelihood}}\,\underbrace{\pi_0}_{\text{prior}}}^{f_1(x)\,=\,\mathrm{P}(H_0|\mathrm{x})}\ \underset{H_1}{\overset{H_0}{\gtreqless}}\ \overbrace{p(x|H_1)\,\pi_1}^{f_2(x)\,=\,\mathrm{P}(H_1|\mathrm{x})}$$

+ the posterior, conditional density given x, is a function of x

Example: Computing two discriminant functions



$$P[H0] = \frac{1}{3}$$

$$P[H1] = \frac{2}{3}$$

+ In this example,
  the class precision is H1 for all x.

- MAP Rule Minimizes Expected Classification Error E[R]
- Classification based on Posterior density is optimal!

$$
\begin{aligned}
E[R] = \pi_0 \cdot E[R|H_0] + \pi_1 E[R|H_1] &= \pi_1 \cdot \int_{y_0} f(y|H_1)dy + \pi_0 \cdot \int_{y_1} f(y|H_0)dy \\
&= \pi_1 \cdot \int_{y_0} f(y|H_1)dy + \pi_0 \cdot (1 - \int_{y_0} f(y|H_0)dy) \\
&= \pi_0 + \int_{y_0} \pi_1 \cdot f(y|H_1) - \pi_0 \cdot f(y|H_0)dy
\end{aligned}
$$

Q: How should we set the decision rule for y_0?

for y, if $\pi_1 \cdot f(y|H_1) - \pi_0 \cdot f(y|H_0) < 0$ then detect as H_0
else if $\pi_1 \cdot f(y|H_1) - \pi_0 \cdot f(y|H_0) \geq 0$ then detect as H_1

$$\frac{p(y|H_1)}{P(y|H_0)} \gtreqless \frac{\pi_0}{\pi_1}$$

"MAP rule"

MAP rule is optimal.

It minimizes expected misclassification error.

$$\overbrace{\underbrace{p(x|H_0)}_{\text{Likelihood!}}\underbrace{\pi_0}_{\text{Prior!}}}^{p(H_0|x)} \underset{H_1}{\overset{H_0}{\gtreqless}} \overbrace{p(x|H_1)\,\pi_1}^{p(H_1|x)}$$

Gaussian **D**iscriminant **A**nalysis (**GDA**)
: this is a MAP rule
as assuming the conditional likelihood is Gaussian.

$$\overbrace{\underbrace{p(x|H_0)}_{\text{Gaussian likelihood}}\,\underbrace{\pi_0}_{\text{Prior}}}^{p(H_0|x)} \underset{H_1}{\overset{H_0}{\gtreqless}} \overbrace{p(x|H_1)\,\pi_1}^{p(H_1|x)}$$

Q: Decision Boundaries for the Two Possible Cases

- $\Sigma_0 \neq \Sigma_1$ (two class sample's covariance are not same)
- $\Sigma_0 = \Sigma_1$

$$\frac{1}{\sqrt{2\pi|\Sigma_0|}}\exp -\frac{1}{2}(x-\mu_0)^t\Sigma_0^{-1}(x-\mu_0)\cdot P[\mathcal{H}_0=0] \underset{H_1}{\overset{H_0}{\gtrless}} \frac{1}{\sqrt{2\pi|\Sigma_1|}}\exp -\frac{1}{2}(x-\mu_1)^t\Sigma_1^{-1}(x-\mu_1)\cdot P[\mathcal{H}_0=1]$$

$$\Updownarrow$$

$$\ln P[\mathcal{H}_0]+\ln\frac{1}{\sqrt{2\pi|\Sigma_0|}}-\frac{1}{2}(x-\mu_0)^t\Sigma_0^{-1}(x-\mu_0) \gtrless \ln P[\mathcal{H}_1]+\ln\frac{1}{\sqrt{2\pi|\Sigma_1|}}-\frac{1}{2}(x-\mu_1)^t\Sigma_1^{-1}(x-\mu_1)$$

- Two Quadratic discriminant functions!
- (how they are shaped over the features space?)
- Quadratic decision boundary!

Q: Decision Boundaries for the Two Possible Cases

- $\Sigma_0 = \Sigma_1$

$$\frac{1}{\sqrt{2\pi|\Sigma_0|}}\exp -\frac{1}{2}(x-\mu_0)^t\Sigma_0^{-1}(x-\mu_0)\cdot P[\mathcal{H}_0=0] \underset{H_1}{\overset{H_0}{\gtrless}} \frac{1}{\sqrt{2\pi|\Sigma_1|}}\exp -\frac{1}{2}(x-\mu_1)^t\Sigma_1^{-1}(x-\mu_1)\cdot P[\mathcal{H}_0=1]$$

$$\Updownarrow$$

$$\ln P[\mathcal{H}_0] + \ln\frac{1}{\sqrt{2\pi|\Sigma_0|}} - \frac{1}{2}(x-\mu_0)^t\Sigma_0^{-1}(x-\mu_0) \gtrless \ln P[\mathcal{H}_1] + \ln\frac{1}{\sqrt{2\pi|\Sigma_1|}} - \frac{1}{2}(x-\mu_1)^t\Sigma_1^{-1}(x-\mu_1)$$

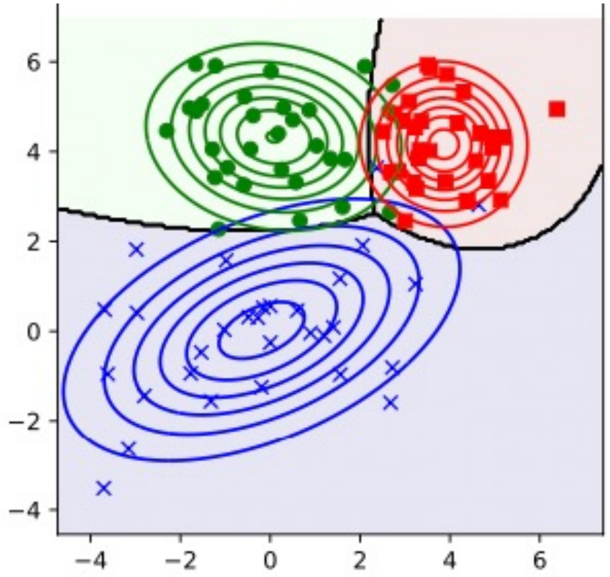
$$\Updownarrow$$

$$\overbrace{\ln P[\mathcal{H}_0]}^{\text{prior}} + \overbrace{\mu_0^t\Sigma_0^{-1}x - \frac{1}{2}\mu_0^t\Sigma_0^{-1}\mu_0^t}^{\text{likelihood}} \gtrless \ln P[\mathcal{H}_1] + \mu_1^t\Sigma_1^{-1}x - \frac{1}{2}\mu_1^t\Sigma_1^{-1}\mu_1^t$$

- Two Linear discriminant functions!
- Linear decision boundary!

From Murphy Figure 9.2

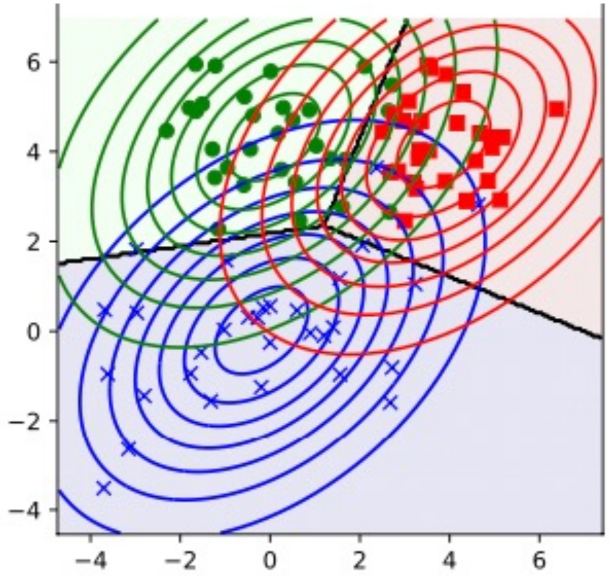
[Quadratic Discriminant Analysis]



$\Sigma_0 \neq \Sigma_1$

(Unconstrained Covariance)

[Linear Discriminant Analysis]



$\Sigma_0 = \Sigma_1$

(Tied Covariance)

QDA becomes LDA as assuming tied Covariance.

case 1]

- scalar feature
- $\sigma_0 = \sigma_1 = \sigma$
- $P[\mathcal{H}_0] = P[\mathcal{H}_1]$

$$\ln P[\mathcal{H}_0] + \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(x-\mu_0)^2 \underset{H_1}{\overset{H_0}{\gtrless}} \ln P[\mathcal{H}_1] + \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(x-\mu_1)^2$$

$$\Updownarrow$$

$$\ln P[\mathcal{H}_0] + \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(x-\mu_0)^2 \gtrless \ln P[\mathcal{H}_1] + \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(x-\mu_1)^2$$

$$\frac{1}{\sigma^2}x\mu_0 - \frac{1}{2\sigma^2}\mu_0^2 \gtrless \frac{1}{\sigma^2}x\mu_1 - \frac{1}{2\sigma^2}\mu_1^2$$

$$x(\mu_0 - \mu_1) \gtrless \frac{1}{2}(\mu_0^2 - \mu_1^2)$$

$$\boxed{x \gtrless \frac{1}{2}(\mu_0 + \mu_1)} \longrightarrow$$

WLOG if $(\mu_0 > \mu_1)$
Binary classification
decision rule!

case 2]

- feature vector
- $\Sigma_0 = \Sigma_1 = \sigma I$, isotropic
- $P[\mathcal{H}_0] = P[\mathcal{H}_1]$

$$\ln P[\mathcal{H}_0] + \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(x-\mu_0)^2 \overset{H_0}{\underset{H_1}{\gtrless}} \ln P[\mathcal{H}_1] + \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(x-\mu_1)^2$$

↕

$$\ln P[\mathcal{H}_0] + \ln \frac{1}{\sqrt{2\pi\sigma^N}} + -\frac{1}{2\sigma^2}(x-\mu_0)^t(x-\mu_0) \gtrless \ln P[\mathcal{H}_1] + \ln \frac{1}{\sqrt{2\pi\sigma^N}} + -\frac{1}{2\sigma^2}(x-\mu_1)^t(x-\mu_1)$$

$$\frac{1}{\sigma^2}\mu_0^t x - \frac{1}{2\sigma^2}\mu_0^t\mu_0 \gtrless \frac{1}{\sigma^2}\mu_1^t x - \frac{1}{2\sigma^2}\mu_1^t\mu_1$$

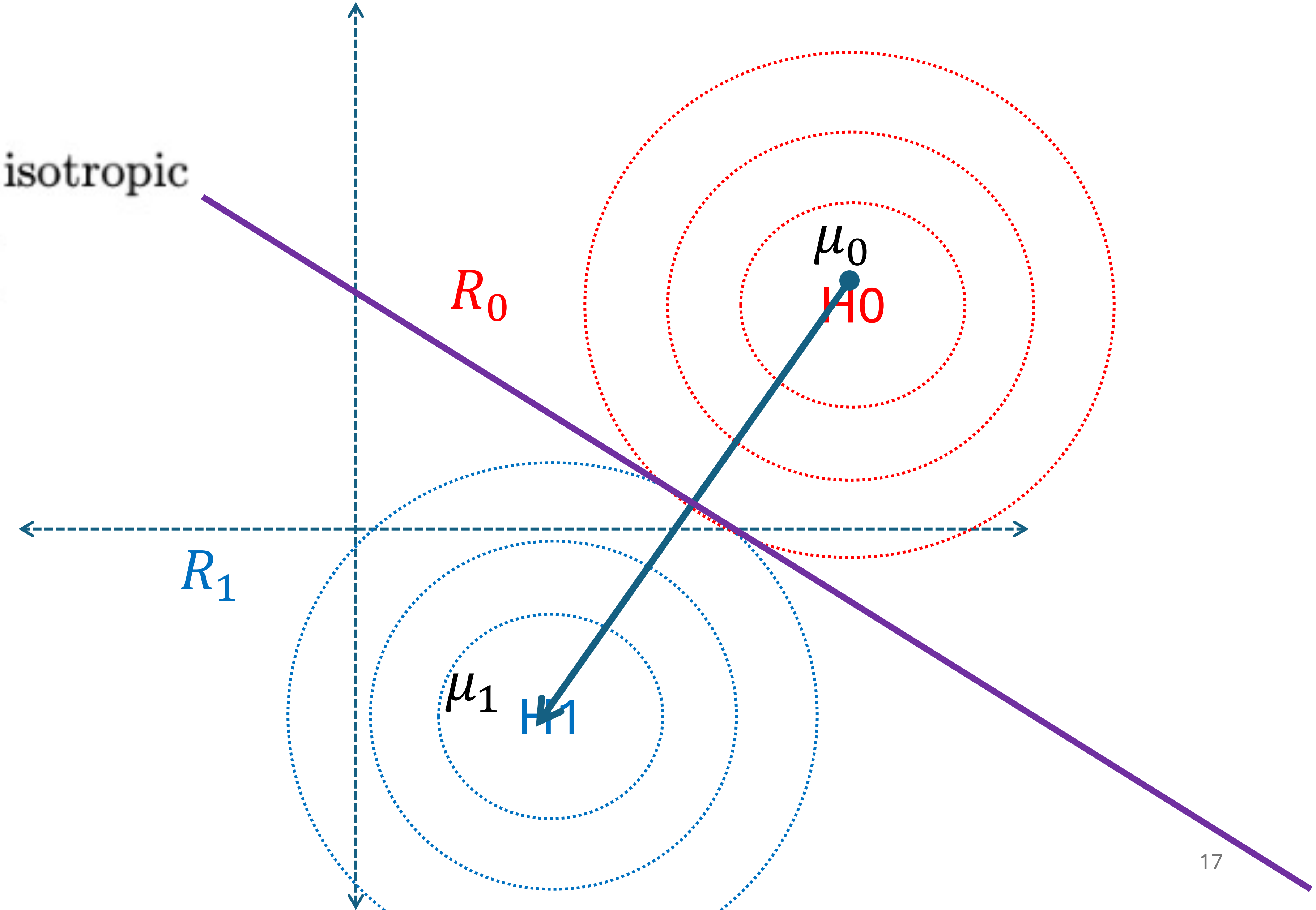
$$\boxed{(\mu_0-\mu_1)^t x \gtrless \frac{1}{2}(\mu_0-\mu_1)^t(\mu_0+\mu_1)}$$

↓ the projection to (μ_0 - μ_1)

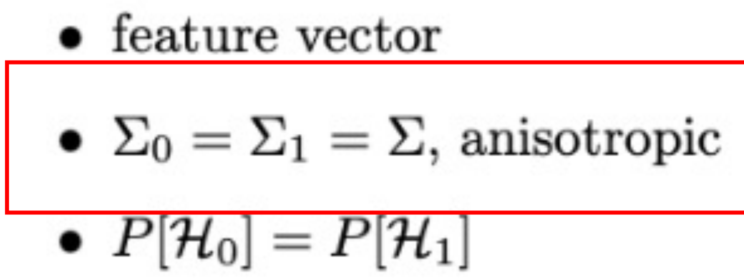
then the decision rule same as the scalar case.

Example) Draw the decision boundaries

- feature vector
- $\Sigma_0 = \Sigma_1 = \sigma I$, isotropic
- $P[\mathcal{H}_0] = P[\mathcal{H}_1]$

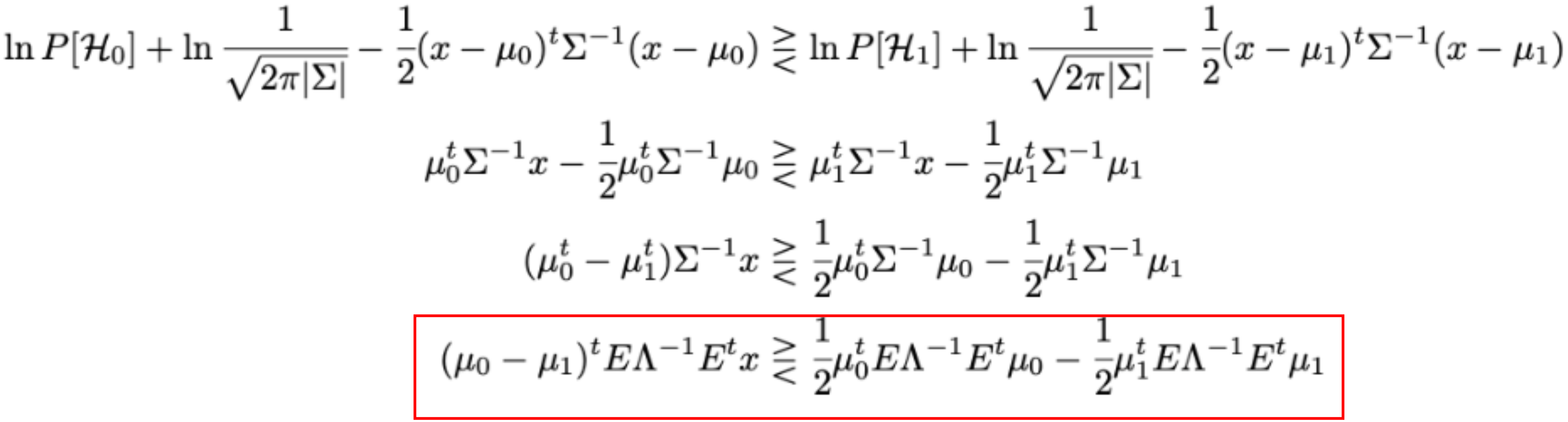


case 3]



- feature vector
- $\Sigma_0 = \Sigma_1 = \Sigma$, anisotropic
- $P[\mathcal{H}_0] = P[\mathcal{H}_1]$

Q: projection to $(\mu'_0 - \mu'_1)$ then the decision rule same as the scalar case?



$$\ln P[\mathcal{H}_0] + \ln \frac{1}{\sqrt{2\pi|\Sigma|}} - \frac{1}{2}(x-\mu_0)^t\Sigma^{-1}(x-\mu_0) \underset{<}{\geq} \ln P[\mathcal{H}_1] + \ln \frac{1}{\sqrt{2\pi|\Sigma|}} - \frac{1}{2}(x-\mu_1)^t\Sigma^{-1}(x-\mu_1)$$

$$\mu_0^t\Sigma^{-1}x - \frac{1}{2}\mu_0^t\Sigma^{-1}\mu_0 \underset{<}{\geq} \mu_1^t\Sigma^{-1}x - \frac{1}{2}\mu_1^t\Sigma^{-1}\mu_1$$

$$(\mu_0^t - \mu_1^t)\Sigma^{-1}x \underset{<}{\geq} \frac{1}{2}\mu_0^t\Sigma^{-1}\mu_0 - \frac{1}{2}\mu_1^t\Sigma^{-1}\mu_1$$

$$\boxed{(\mu_0 - \mu_1)^t E\Lambda^{-1}E^t x \underset{<}{\geq} \frac{1}{2}\mu_0^t E\Lambda^{-1}E^t\mu_0 - \frac{1}{2}\mu_1^t E\Lambda^{-1}E^t\mu_1}$$

$$\boxed{(\mu_0-\mu_1)^t E\Lambda^{-1}E^t x \gtrless \frac{1}{2}\mu_0^t E\Lambda^{-1}E^t\mu_0 - \frac{1}{2}\mu_1^t E\Lambda^{-1}E^t\mu_1}$$

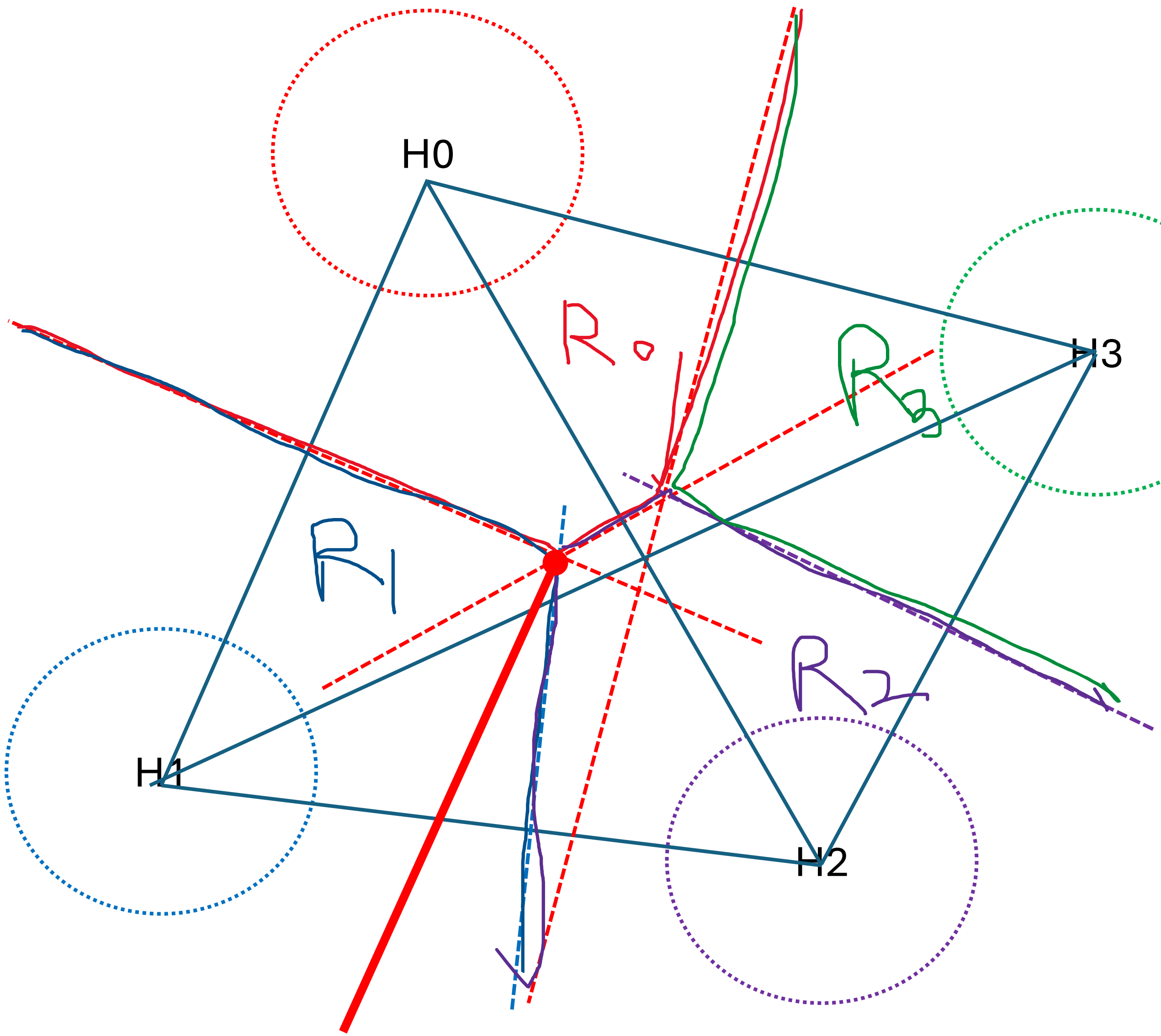
$$(\mu_0-\mu_1)^t E\Lambda^{-1}E^t x \gtrless \frac{1}{2}\mu_0^t E\Lambda^{-1}E^t\mu_0 - \frac{1}{2}\mu_1^t E\Lambda^{-1}E^t\mu_1$$

$$[\Lambda^{-1/2}E^t(\mu_0-\mu_1)]^t[\Lambda^{-1/2}E^t]x \gtrless \frac{1}{2}[\Lambda^{-1/2}E^t(\mu_0-\mu_1)]^t[\Lambda^{-1/2}E^t](\mu_0+\mu_1)$$

$$\underbrace{[\Lambda^{-1/2}E^t(\mu_0-\mu_1)]^t}_{\text{+projection to }(\mu_0\ -\mu_1)}\underbrace{[\Lambda^{\frac{-1}{2}}E^t]}_{\text{+whitening (even though no centering)}}x \gtrless \frac{1}{2}(\mu_0-\mu_1)^t E\Lambda^{-1}E^t(\mu_0+\mu_1)$$

case 4] multi-class cases

- feature vector
- $\Sigma_0 = \Sigma_1 = \Sigma_3 = \Sigma_4 = \sigma I$, isotropic
- $P[\mathcal{H}_0] = P[\mathcal{H}_1] = P[\mathcal{H}_2] = P[\mathcal{H}_3]$



+the lines meet at one point (circumcenter)

Note that
the MAP rule minimizes the mis-classification Error!
So, it is an optimal solution when we know the likelihood and prior.

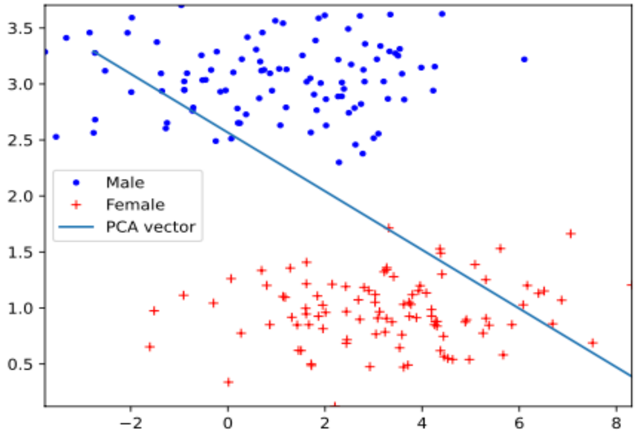
One way to view the linear classification is in terms of dimensionality reduction.
And compare the values in the new projection space.
Which directional projection will be beneficial to classification?

$$\boxed{(\mu_0 - \mu_1)^t x \underset{<}{\geq} \frac{1}{2}(\mu_0 - \mu_1)^t(\mu_0 + \mu_1)}$$

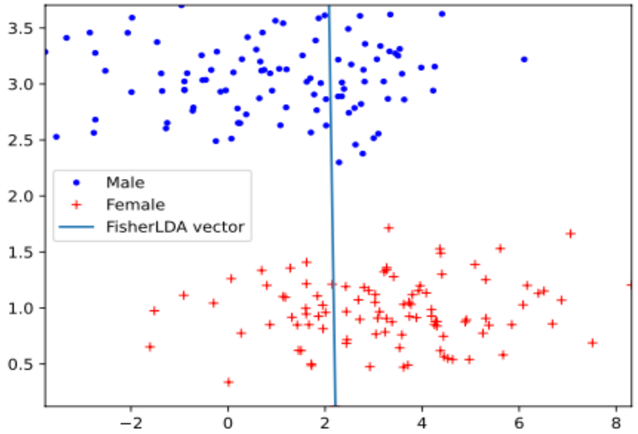
$$\boxed{(\mu_0 - \mu_1)^t E\Lambda^{-1}E^t x \underset{<}{\geq} \frac{1}{2}\mu_0^t E\Lambda^{-1}E^t \mu_0 - \frac{1}{2}\mu_1^t E\Lambda^{-1}E^t \mu_1}$$

+ projection !
+ compare projected values with a threshold!

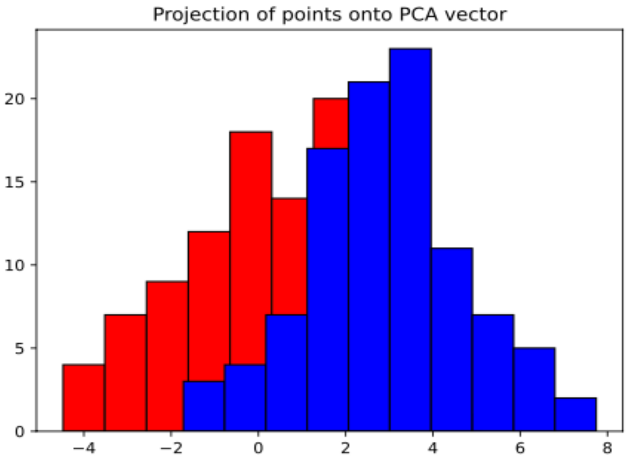
One way to view a linear classification is in terms of dimensionality reduction. Which directional projection will be beneficial to classification?



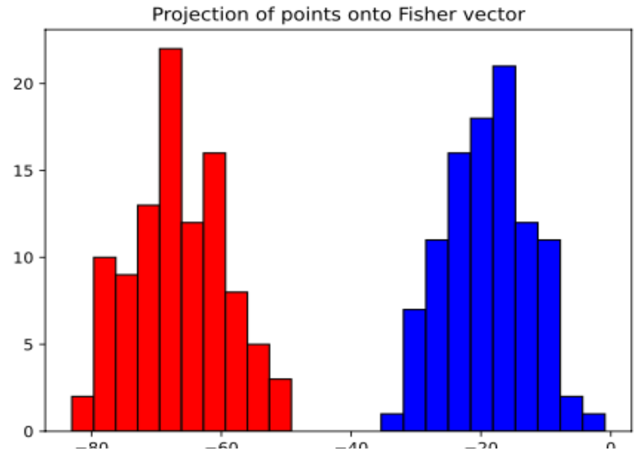
(a)



(b)



PCA?



Fisher’s Linear Discriminant Analysis

Fisher’s Discriminant Analysis

It defines an ideal projection for classification.

Fisher’s Discriminant Analysis for binary classes

$$y = \boxed{\vec{w}^{t}} \cdot \vec{x}$$

We want to define a projection ($\vec{w}$)
which gives an ideal space for classification.

- The distance between the means of the projections: $||\mu_{y0} - \mu_{y1}||$
- The variance of the projections: ${\sigma_{y0}}^2 + {\sigma_{y1}}^2$

Fisher’s Discriminant Analysis

$$y = \vec{w}^t \cdot \vec{x}$$

We want to define a projection ($\vec{w}$)
which gives an ideal space for classification.

- The distance between the means of the projections: $||\mu_{y0} - \mu_{y1}||$

- The variance of the projections: $\sigma_{y0}{}^2 + \sigma_{y1}{}^2$

- $\vec{w} = argmax_{\vec{w}}$J(w)

$$J(\vec{w}) = \frac{\vec{w}^t(\mu_{x0} - \mu_{x1})(\mu_{x0} - \mu_{x1})^t\vec{w}}{\vec{w}^t\{\sum_{c0}(x_n - \mu_{x0})(x_n - \mu_{x0})^t + \sum_{c1}(x_n - \mu_{x1})(x_n - \mu_{x1})^t\}\vec{w}}$$

- Finding the optimal point $\overrightarrow{W}$ by $\nabla J(\overrightarrow{W}) = 0$

$$J(\vec{w}) = \frac{\vec{w}^t \overbrace{(\mu_{x0} - \mu_{x1})(\mu_{x0} - \mu_{x1})^t}^{S_B} \vec{w}}{\vec{w}^t \underbrace{\{\sum_{c0}(x_n - \mu_{x0})(x_n - \mu_{x0})^t + \sum_{c1}(x_n - \mu_{x1})(x_n - \mu_{x1})^t\}}_{S_W} \vec{w}}$$

$$J(\vec{w}) = \frac{\vec{w}^t S_B \vec{w}}{\vec{w}^t S_W \vec{w}}$$

$$\nabla J(\vec{w}) = \frac{(\vec{w}^t S_W \vec{w})(S_B \vec{w}) - (S_W \vec{w})(\vec{w}^t S_B \vec{w})0}{||\vec{w}^t S_W \vec{w}||^2} = 0$$

$$\leftrightarrow (\underline{\vec{w}^t S_W \vec{w}})(S_B \vec{w}) = (S_W \vec{w})(\underline{\vec{w}^t S_B \vec{w}})$$

$$(S_W \vec{w}) = \alpha(\mu_{x0} - \mu_{x1})$$

$$\vec{w} = \alpha S_W^{-1}(\mu_{x0} - \mu_{x1})$$

The relation between Fisher's Discriminant Projection & Gaussian Discriminant Analysis

Fisher Discriminant Projection

$$(S_W \vec{w}) = \alpha(\mu_{x0} - \mu_{x1})$$

$$\vec{w} = \alpha \boxed{S_W^{-1}} (\mu_{x0} - \mu_{x1})$$

When the two covariance matrices are equivalent, Fisher's projection is same thing with the anisotropic case in GDA.

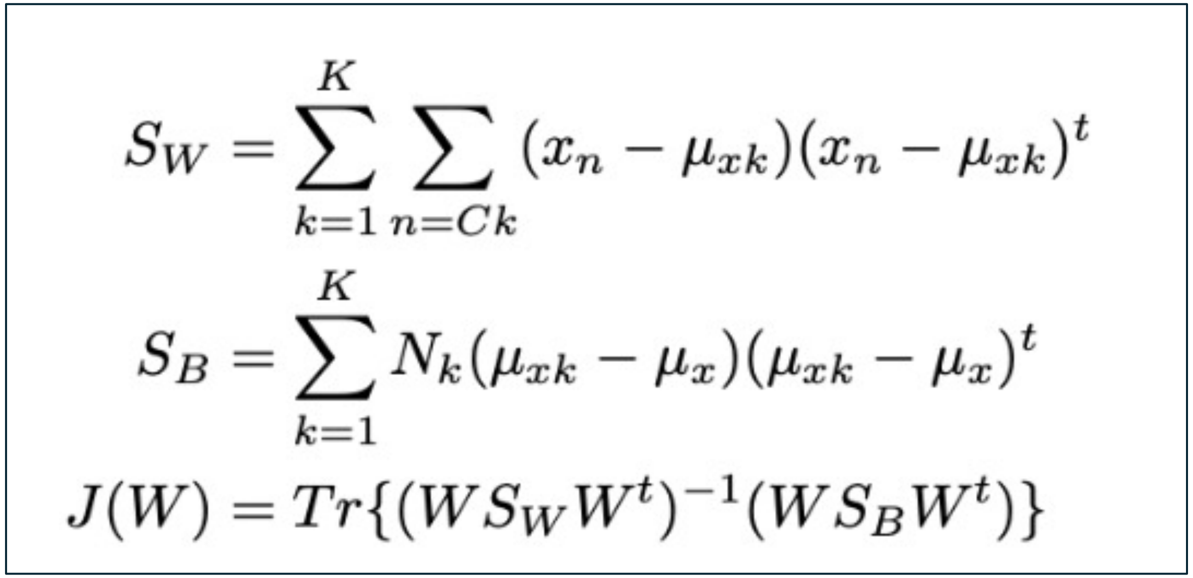
Gaussian Discriminant Analysis

- Whitening
- Projection to $(\mu_0 - \ \mu_1)$

1. $\Lambda^{-1/2} E^t x$
2. $(\Lambda^{-1/2} E^t \mu_{x0} - \Lambda^{-1/2} E^t \mu_{x1})^t \Lambda^{-1/2} E^t x$
3. $(\mu_{x0}^t E \Lambda^{-1/2} - \mu_{x1}^t E \Lambda^{-1/2}) \Lambda^{-1/2} E^t x$
4. $(\mu_{x0} - \mu_{x1})^t \boxed{E \Lambda^{-1} E^t} x$

Fisher's Discriminant Analysis for Multiple K Classes

$$y = W \cdot \vec{x}$$ (the size of W is D' x D)



$$S_W = \sum_{k=1}^{K} \sum_{n=Ck} (x_n - \mu_{xk})(x_n - \mu_{xk})^t$$

$$S_B = \sum_{k=1}^{K} N_k(\mu_{xk} - \mu_x)(\mu_{xk} - \mu_x)^t$$

$$J(W) = Tr\{(WS_W W^t)^{-1}(WS_B W^t)\}$$

Q: What is the rank of S_B?

The optimal projection matrix W*

: eigenvectors of $(S_W^{-1} S_B)$ *$that\ corresponds\ to\ the\ D' largest\ eigenvalues$*.

+ please note that $the\ rank\ of\ S_B$ is at most K-1 (where K is the number of classes). That means the number of positive eigenvalues are K-1 so the dimension of D' (the dim of projection space) is restricted by K-1.

Fisher’s linear discriminant does not provide discriminant functions.
It provides a specific choice of direction for projection.

Hence,
the Fisher’s projected data can subsequently used
to construct a discriminant such as GDA.

Generative vs. Discriminative Modeling in Classification

Generative Classification

$$\underbrace{\arg\max_{k} P[C_k|x]} \propto P[x, C_k] = \underbrace{P[x|C_k]P[x]}_{P[C_k\,|x] \,\propto\, P[x,\, C_k]}$$

In the prediction/inference stage,
we compare the posterior
(discriminant functions)

In the training stage,
we learn the likelihood and prior,
so posterior (joint density!)

GDA is generative Modeling (Example1)

As we train the discriminant functions, We estimate prior, mean, covariance for each class to define posterior.

- feature vector
- $\Sigma_0 = \Sigma_1 = \Sigma$, anisotropic
- $P[\mathcal{H}_0] = P[\mathcal{H}_1]$

$$\ln P[\mathcal{H}_0] + \ln\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(x-\mu_0)^2 \gtrless \ln P[\mathcal{H}_1] + \ln\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(x-\mu_1)^2$$

$$\Updownarrow$$

$$\ln P[\mathcal{H}_0] + \ln\frac{1}{\sqrt{2\pi|\Sigma|}} - \frac{1}{2}(x-\mu_0)^t\Sigma^{-1}(x-\mu_0) \gtrless \ln P[\mathcal{H}_1] + \ln\frac{1}{\sqrt{2\pi|\Sigma|}} - \frac{1}{2}(x-\mu_1)^t\Sigma^{-1}(x-\mu_1)$$

$$\mu_0^t\Sigma^{-1}x - \frac{1}{2}\mu_0^t\Sigma^{-1}\mu_0 \gtrless \mu_1^t\Sigma^{-1}x - \frac{1}{2}\mu_1^t\Sigma^{-1}\mu_1$$

$$(\mu_0^t - \mu_1^t)\Sigma^{-1}x \gtrless \frac{1}{2}\mu_0^t\Sigma^{-1}\mu_0 - \frac{1}{2}\mu_1^t\Sigma^{-1}\mu_1$$

$$\boxed{(\mu_0 - \mu_1)^t E\Lambda^{-1}E^t x \gtrless \frac{1}{2}\mu_0^t E\Lambda^{-1}E^t\mu_0 - \frac{1}{2}\mu_1^t E\Lambda^{-1}E^t\mu_1}$$

$$\downarrow$$

Naïve Bayes is one example of GDA (Example2)
Generative Modeling

- $f_0(g,b)$

$$P[D=+|G=g,B=b]=\frac{P[D=+,G=g,B=b]}{P[G=g,B=b]}=\frac{P[G=g|D=+]\cdot P[B=b|D=+]\cdot P[D=+]}{P[G=g,B=b]}$$

$$P[D=-|G=g,B=b]=\frac{P[D=-,G=g,B=b]}{P[G=g,B=b]}=\frac{P[G=g|D=-]\cdot P[B=b|D=-]\cdot P[D=-]}{P[G=g,B=b]}$$

- $f_1(g,b)$

- $f(g,b|D+)$
- $f(g,b|D-)$
- $f(D+)$
- $f(D-)$

In HW1, we estimated these statistics so able to find the two discriminant functions (posteriors).

Discriminative Classification

Discriminative Classification

$$\arg\max_{k} P[C_k|x] \propto P[x, C_k] = P[x|C_k]P[x]$$

In GDA (generative modeling),
we found that the posterior becomes the linear function of X (feature/ data).

$$\boxed{(\mu_0 - \mu_1)^t E\Lambda^{-1}E^t x \gtrless \frac{1}{2}\mu_0^t E\Lambda^{-1}E^t\mu_0 - \frac{1}{2}\mu_1^t E\Lambda^{-1}E^t\mu_1}$$

Discriminative Classification

: can directly model the posterior with a linear function of feature map?
and directly learn the linear function?
without estimating likelihood / prior?

Discriminative Modeling $P[C_1|x]$ using logistic sigmoid function.
Let’s focus on binary classification today.

$$P[C_1|x] = \frac{P[x|C_1]P[C_1]}{P[x|C_1]P[C_1 + P[x|C_0]P[C_0]}$$

$$P[C_1|x] = \frac{1}{1 + \frac{P[x|C_0]P[C_0]}{P[x|C_1]P[C_1]}}$$

$$P[C_1|x] = \frac{1}{1 + \exp\left(\ln \frac{P[x|C_0]P[C_0]}{P[x|C_1]P[C_1]}\right)}$$

$$P[C_1|x] = \frac{1}{1 + \exp\left(-\ln \frac{\boxed{P[x|C_1]P[C_1]}}{P[x|C_0]P[C_0]}\right)} \longrightarrow$$

The natural log would linearly combine the two posteriors which are linear forms.

Discriminative Modeling $P[C_1|x]$ using logistic sigmoid function.

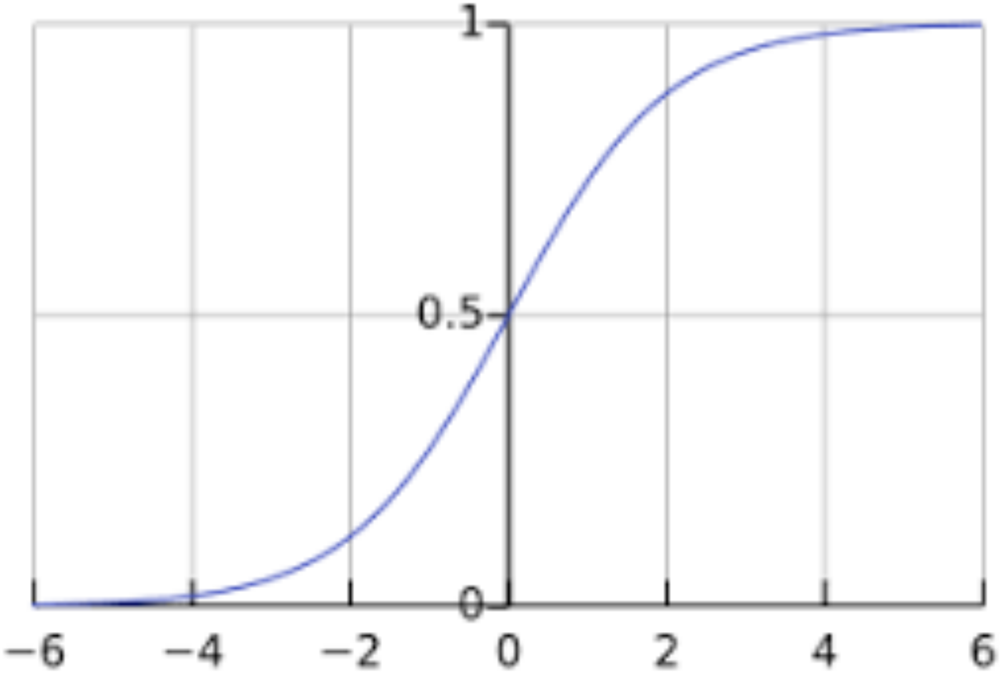
$$P[C_1|x] = \frac{P[x|C_1]P[C_1]}{P[x|C_1]P[C_1 + P[x|C_0]P[C_0]}$$

$$P[C_1|x] = \frac{1}{1 + \frac{P[x|C_0]P[C_0]}{P[x|C_1]P[C_1]}}$$

$$P[C_1|x] = \frac{1}{1 + \exp\,(\ln \frac{P[x|C_0]P[C_0]}{P[x|C_1]P[C_1]})}$$

$$P[C_1|x] = \frac{1}{1 + \exp\,(-\ln \boxed{\frac{P[x|C_1]P[C_1]}{P[x|C_0]P[C_0]})}} \longrightarrow W^t\, x \text{ or } W^t\, \phi(x)$$

Discriminative Modeling $P[C_1|x]$ using logistic sigmoid function.

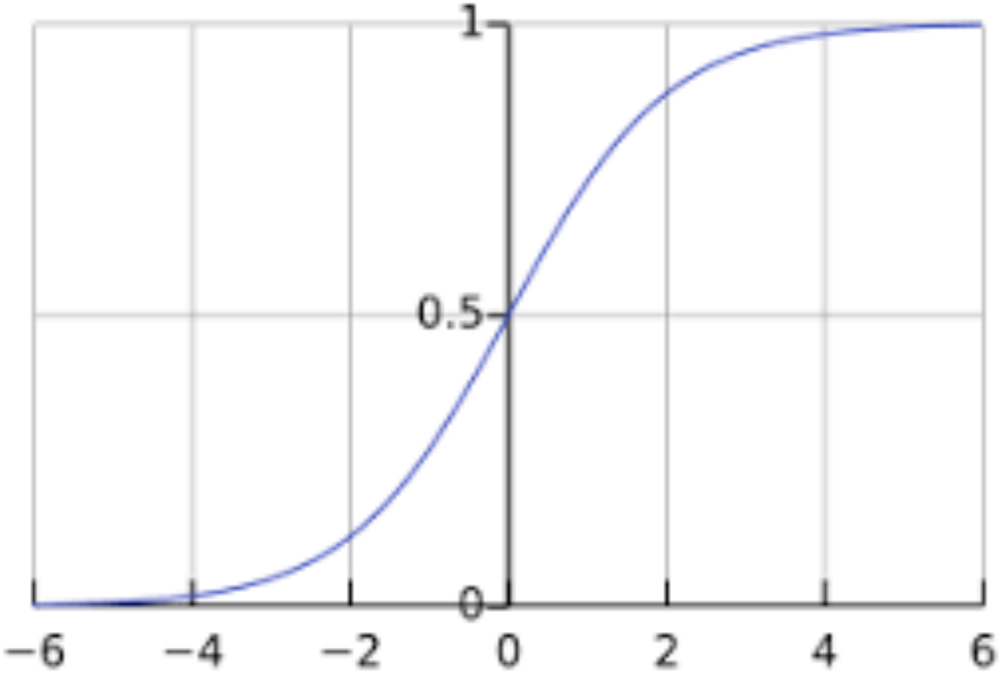


$$\sigma(x) = \frac{1}{1 + exp^{-x}}$$

$$P[C_1|\vec{x}] = \frac{1}{1 + \exp\left(-\vec{w}^t\vec{x}\right)}$$

$$P[C_0|\vec{x}] = \frac{\exp\left(-\vec{w}^t\vec{x}\right)}{1 + \exp\left(-\vec{w}^t\vec{x}\right)}$$

Discriminative Modeling $P[C_1|x]$ using logistic sigmoid function.



$$P[C_1|\vec{x}] = \frac{1}{1 + \exp\left(-\vec{w}^t\vec{x}\right)}$$

$$P[C_0|\vec{x}] = \frac{\exp\left(-\vec{w}^t\vec{x}\right)}{1 + \exp\left(-\vec{w}^t\vec{x}\right)}$$

We need to estimate W, but how?
Estimation rule!

$$\sigma(x) = \frac{1}{1 + exp^{-x}}$$

Bayes Rule again to Estimate W

$$P(w|D) = \frac{p(w, D)}{P(D)} = \frac{p(D|w)p(w)}{p(D)}$$

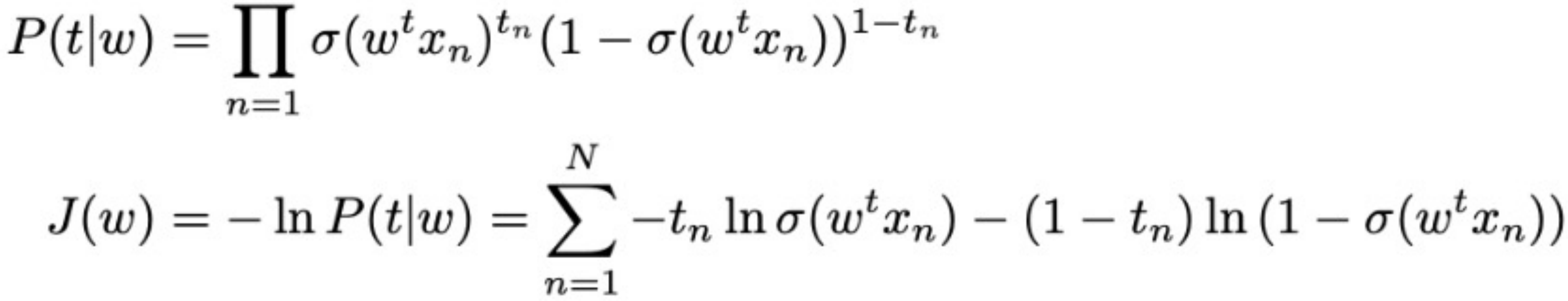
Today,
We are going to focus on MLE.

$$P(w|D) = \frac{p(w,D)}{P(D)} = \frac{\boxed{p(D|w)}p(w)}{p(D)}$$

MLE: Logistic Regression

- Suppose we have a data set $\{x_n, t_n\}$ $where\ t_n \in \{0,1\}$ $and\ n = \{1,2, \ldots, N\}$
- The likelihood function can be written as

$$\mathrm{P}[T_n = t] = p^t(1-p)^{1-t} \qquad (T_n\ is\ Benoulli\ R.V)$$



$$P(t|w) = \prod_{n=1} \sigma(w^t x_n)^{t_n}(1-\sigma(w^t x_n))^{1-t_n}$$

$$J(w) = -\ln P(t|w) = \sum_{n=1}^{N} -t_n \ln \sigma(w^t x_n) - (1-t_n)\ln(1-\sigma(w^t x_n))$$

Q: Then how can we find W*?

MLE: Logistic Regression

Finding the optimal point $\overrightarrow{W}$ by $\nabla J(\overrightarrow{W}) = 0$??

$$\nabla_w J(w) = \sum_{n=1}^{N} -t_n \frac{\sigma(w^t x_n)(1-\sigma(w^t x_n))}{\sigma(w^t x_n)} x_n - (1-t_n) \frac{-\sigma(w^t x_n)(-1+\sigma(w^t x_n))}{1-\sigma(w^t x_n)} x_n$$

$$= \sum_{n=1}^{N} \{-t_n(1-\sigma(w^t x_n)) - (1-t_n)(-\sigma(w^t x_n)\} x_n$$

$$\boxed{= \sum_{n=1}^{N} (\sigma(w^t x_n) - t_n) x_n}$$

no closed form for $\nabla_w J(w) = 0$

Q: Then how can we find W*?

MLE: Logistic Regression
Gradient Descent Algorithm

$$w_{i+1} = w_i - \eta \nabla J(w)$$

$$w_{i+1} = w_i - \eta \sum_{n=1}^{N} (\sigma(w^t x_n) - t_n) x_n$$

- $gradient\ gives\ the\ steepest\ direction$!

  ex) $f(x, y) = x^2 + y^2$

Logistic regression has a global minimum,
so any initial point will converge to the optimal $solution\ W*$
as we have a proper step size.

In the next class

- MAP approach for Logistic Regression + adding regularization
- Multiclass Logistic Regression + softmax approach
- Metrics for Classification+ ROC curve
- Move to the topic perceptron