Lecture #3 - Classification (Ch 4)

$Y = f(X) + \varepsilon, \quad E(\varepsilon) = 0$ } "the model"

$Y \in \mathbb{R}$

In lecture #2, we "assumed" that $f(X) = \beta_0 + \beta_1 X$

OR $f(X) = \beta_0 + \beta_1 \sin(X) + \beta_2 \cos(X)$

Today: (Ch 4)

Next week: (Ch 5, cross-validation; Ronan)

$Y \in \{A, B, O, AB\}$ ⟹ Y is qualitative.

If X is qualitative

How about: $Y = \begin{cases} 1 & \text{if } A \\ 2 & \text{if } B \\ 3 & \text{if } O \\ 4 & \text{if } AB \end{cases}$?

then $Y = f(X) + \varepsilon$!

eg: $Y = \beta_0 + \beta_1 X + \varepsilon$

then a $\frac{1}{\beta_1}$ - unit increase in X is associd w/ blood type

$A \Rightarrow B$, or $B \Rightarrow O$,

or $O \Rightarrow AB$ . .

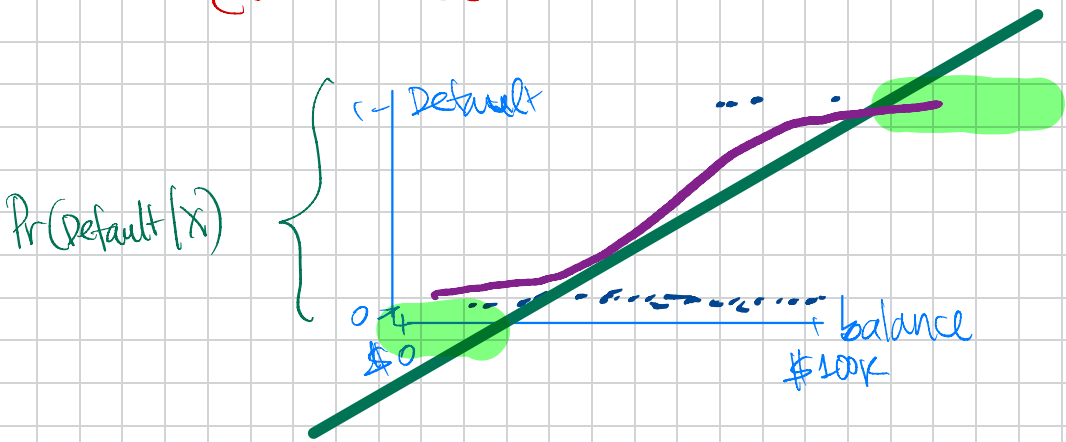
Switching to $Y = \begin{cases} 1 & \text{if } B \\ 2 & \text{if } O \\ 3 & \text{if } AB \\ 4 & \text{if } A \end{cases}$

Note: if Y is binary (eg. default vs. non-default), then OK to code $Y = \begin{cases} 1 & \text{if default} \\ 0 & \text{if not default} \end{cases}$ and do regression.

if all you care about is prediction.

$$Y = \begin{cases} 1 & \text{if default} \\ 0 & \text{else} \end{cases}$$

X = balance



Logistic Regression.

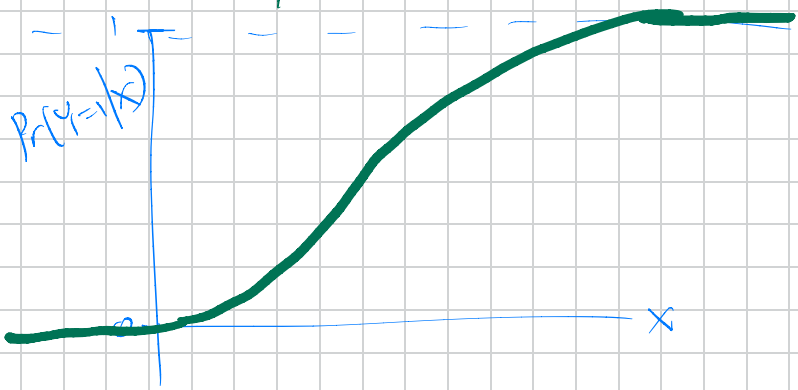
$$Y = \begin{cases} 1 & \text{if default} \\ 0 & \text{if not} \end{cases}$$

X = balance

Our goal: model $\Pr(Y=1 \mid X=x) = \dfrac{\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)}$

$$\exp(a) = e^a = 2.7828^a$$

If $\beta_1 > 0$:



Define $p(X) := \Pr(Y=1 \mid X)$

$$1-p(X) = \Pr(Y=0 \mid X)$$

$$= \frac{1}{1+\exp(\beta_0+\beta_1 X)}$$

$$Pr(Y=1|X) = p(X) = \frac{\exp(\beta_0 + \beta_1 X)}{1+\exp(\beta_0+\beta_1 X)}$$

$$p(X)\left(1+\exp(\beta_0+\beta_1 X)\right) = \exp(\beta_0+\beta_1 X)$$

$$p(X) = \exp(\beta_0+\beta_1 X)\left[1-p(X)\right]$$

$$\frac{p(X)}{1-p(X)} = \exp(\beta_0+\beta_1 X)$$

$$\beta_0+\beta_1 X = \log\left(\frac{p(X)}{1-p(X)}\right) = \overbrace{\log\left(\frac{Pr(Y=1|X)}{Pr(Y=0|X)}\right)}^{\text{logit}(X)}$$

Q: a one-unit increase in X is assoc'd w/ how much increase in $Pr(Y=1|X)$?

A: $Pr(Y=1|X+1) - Pr(Y=1|X)$

$$= \frac{\exp(\beta_0 + \beta_1(x+1))}{1+\exp(\beta_0+\beta_1(x+1))} - \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$$

Logistic regression:

$$\Pr(Y=1|x) = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$$

$$\Rightarrow Y \sim \text{Bernoulli}\left(\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}\right)$$

Further assume that have n i.i.d. draws of $\{(x_i, y_i)\}_{i=1}^n$ from this model

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \left(\frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}\right)^{y_i} \left(\frac{1}{1+\exp(\beta_0+\beta_1 x_i)}\right)^{1-y_i}$$

$Z \sim \text{Bernoulli}(s)$

$$f(z) = s^z (1-s)^{1-z}$$

Select $\hat{\beta}_0, \hat{\beta}_1$ to maximize

$$\log[L(\beta_0, \beta_1)]$$

$$\hat{Pr}(Y=1 \mid X=x) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)}$$

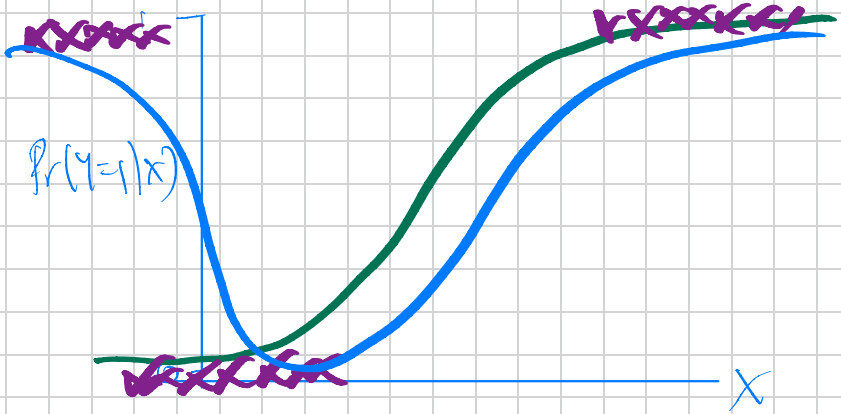
$$\hat{Pr}(Y=0 \mid X=x) = \frac{1}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)}$$

Multiple logistic regression:

$Y \in \{0, 1\}$; $X_1, \ldots, X_p \in \mathbb{R}$

$$Pr(Y=1|X) = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}$$

$$\log\left(\frac{Pr(Y=1|X)}{1-Pr(Y=1|X)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p.$$



$$Pr(Y=1|X) = \frac{\exp(\beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3}{1 + \exp(\beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3)}$$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$$

$$Pr(Y=1|X) = \frac{\exp(\beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 \sin X)}{1 + \exp(\beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 \sin X)}$$

$$\text{logit}\left(\frac{Pr(Y=1|X)}{1-Pr(Y=1|X)}\right) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 \sin X$$

Logistic, more generally:

$$Pr(Y=1|X) = \frac{\exp(f(X))}{1+\exp(f(X))}$$

where $f(\cdot)$ is unknown.

Compare: $Y = f(X) + \varepsilon$

Multinomial logistic regression.

Now, instead of Y taking on 2 possible values, it can take $K \geq 2$ values.

e.g. $Y \in \{A, B, AB, O\}$.

$$Y = \begin{cases} 1 & \text{if } A \\ 2 & \text{if } B \\ 3 & \text{if } AB \\ 4 & \text{if } O \end{cases}$$

$(K-1)(p+1)$

For $k = 1, \ldots, K-1$:

$$\Pr(Y = k \mid X = x) = \frac{\exp(\beta_{k0} + \beta_{k1} x_1 + \ldots + \beta_{kp} x_p)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \beta_{l1} x_1 + \ldots + \beta_{lp} x_p)}$$

and

$$\Pr(Y = K \mid X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \beta_{l1} x_1 + \ldots + \beta_{lp} x_p)}$$

Check: $\sum_{k=1}^{K} \Pr(Y = k \mid X = x) = 1$

Check: if $K=2$,

$$\Pr(Y=1 \mid X=x) = \frac{\exp(\beta_{10} + \beta_{11}x_1 + \cdots + \beta_{1p}x_p)}{1 + \exp(\beta_{10} + \beta_{11}x_1 + \cdots + \beta_{1p}x_p)}$$

$$\Pr(Y=2 \mid X=x) = \frac{1}{1 + \exp(\beta_{10} + \beta_{11}x_1 + \cdots + \beta_{1p}x_p)}$$

In general, [a.k.a "soft max"]

For $k=1,\ldots,K-1$,

$$\Pr(Y=k \mid X=x) = \frac{\exp(f_k(x))}{1 + \sum_{\ell=1}^{K-1} \exp(f_\ell(x))},$$

and

$$\Pr(Y=K \mid X=x) = \frac{1}{1 + \sum_{\ell=1}^{K-1} \exp(f_\ell(x))}$$

Another way to write (multinomial) logistic regression:

$K \cdot (p+1)$

For $k = 1, \ldots, K$:

$$\Pr(Y = k \mid X = x) = \frac{\exp(\beta_{k0} + \beta_{k1} x_1 + \cdots + \beta_{kp} x_p)}{\sum_{\ell=1}^{K} \exp(\beta_{\ell 0} + \beta_{\ell 1} x_1 + \cdots + \beta_{\ell p} x_p)}$$

and

$$\Pr(Y = K \mid X = x) = \frac{1}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{\ell 0} + \beta_{\ell 1} x_1 + \cdots + \beta_{\ell p} x_p)}$$

LINEAR REG:

$$Y = f(X) + \varepsilon, \quad E(\varepsilon) = 0$$

and I think it's reasonable to suppose

$$f(X) \approx \beta_0 + \beta_1 X,$$

and $\varepsilon_1, \ldots, \varepsilon_n$ indep

LOGISTIC REG [$Y \in \{0, 1\}$]

$$\Pr(Y = 1 \mid X) = \frac{\exp(f(X))}{1 + \exp(f(X))}$$

and reasonable $f(X) \approx \beta_0 + \beta_1 X$

and $Y_1, \ldots, Y_n$ indep.

Discriminant Methods. ($K \geq 2$ classes)

Bayes Rule: $\Pr(A|B) = \dfrac{\Pr(B|A)\Pr(A)}{\Pr(B)}$

$$\Pr(Y=k \mid X=x) = \frac{\Pr(X=x \mid Y=k)\Pr(Y=k)}{\Pr(X=x)}$$

($Y=k$ is A, $X=x$ is B)

$$= \frac{\Pr(X=x \mid Y=k)\Pr(Y=k)}{\sum_{\ell=1}^{K} \Pr(X=x \mid Y=\ell)\Pr(Y=\ell)}$$

Goal: use this to develop a classification method.

Notation: $\pi_k := \Pr(Y=k)$

$f_k(x) = \Pr(X=x \mid Y=k)$

$$Pr(Y=k|X=x) = \frac{f_k(x)\,\pi_k}{\sum_{l=1}^{K} f_l(x)\,\pi_l}$$

① est. π_k, $f_k(\cdot)$

② plug in those ests to get an est of $Pr(Y=k|X=x)$

discriminant method

All that remains is to est. π_k and $f_k(\cdot)$.

To estimate π_k, use $\hat{\pi}_k = \frac{\sum_{i=1}^{n} 1_{\{y_i=k\}}}{n}$

How to estimate $f_k(\cdot) : \mathbb{R}^p \to \mathbb{R}$?

TODAY

- Option 1: linear discriminant analysis
- Option 2: naive Bayes

Linear discriminant analysis:

We will assume that within the k^{th} class ($k = 1, \ldots, K$),

$$X \sim N_p(\mu_k, \Sigma)$$

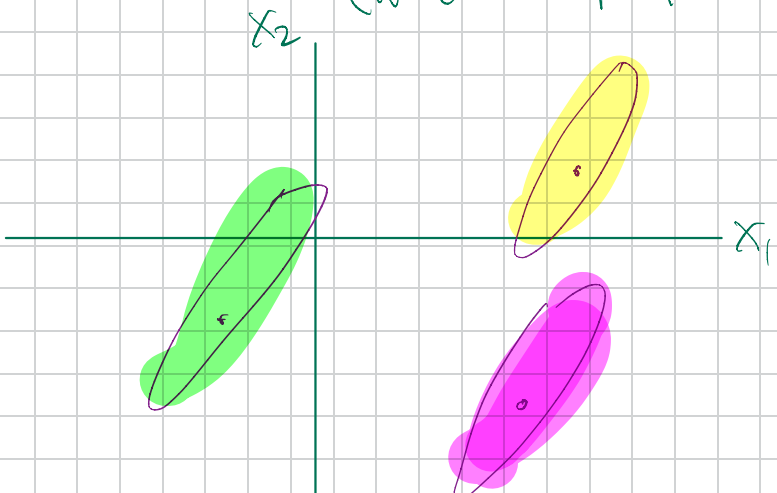
(Σ: $p \times p$; μ_k: p-vector)

X is MVN w/ class-specific mean, and a common VARIANCE

Then,

$$f_k(x) = \underbrace{\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}}_{C} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)\right)$$

(where $|\Sigma| = \det(\Sigma)$)



So,

$$\Pr(Y=k \mid X=x) = \frac{f_k(x)\,\pi_k}{\sum_{l=1}^{K} f_l(x)\,\pi_l}$$

$$= \frac{\hat{f}_k(x)\,\hat{\pi}_k}{\text{denom}(x)}$$

$$= \frac{c \exp\{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)\} \cdot \pi_k}{\text{denom}(x)}$$

denom(x) ← same for all $k = 1, \ldots, K$

$$\propto \exp\{-\tfrac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)\} \cdot \hat{\pi}_k$$

- will assign x to the class for which $\Pr(Y=k \mid X=x)$ largest

To est. μ_k:

$\hat{\mu}_k$ = sample mean of the obs. in the k^{th} class.

$\hat{\Sigma}$ = "pooled" est of Σ.

We assign obs. x to the class for which

$$\Pr(Y=k \mid X=x)$$

$$\log[C^A \cdot B] = A + \log(B)$$

$$\propto \exp\left\{-\tfrac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)\right\} \cdot \hat{\pi}_k$$

is largest.

i.e. we will assign x to class for which

$$-\tfrac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k) + \log(\hat{\pi}_k)$$

is largest,

$$= -\tfrac{1}{2} x^T \Sigma^{-1} x + x^T \Sigma^{-1} \mu_k - \tfrac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\hat{\pi}_k)$$

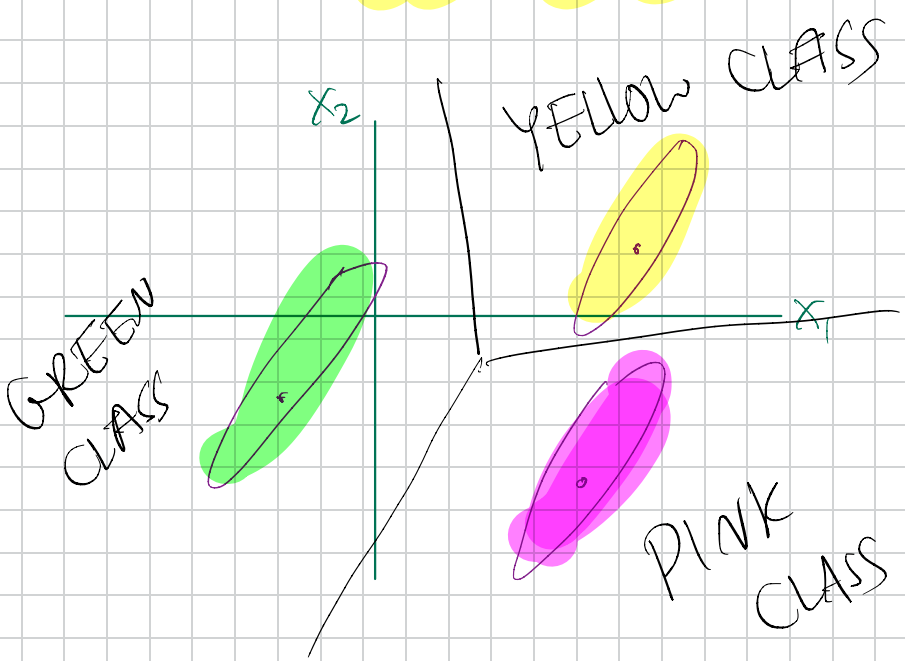
$$= \underbrace{\left(-\tfrac{1}{2} x^T \Sigma^{-1} x\right)}_{c(x)} - \underbrace{\left(\tfrac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\hat{\pi}_k)\right)}_{d_k} + x^T (\Sigma^{-1} \mu_k)$$

$$= c(x) + d_k + x^T(\Sigma^{-1}\mu_k)$$

$\Rightarrow$ So, I'm assigning x to the class for which

$$\cancel{c(x)} + d_k + x^T(\Sigma^{-1}\mu_k)$$

is largest.



Linear disc. analysis:

(1) recalling Bayes:

$$Pr(A|B) = \frac{Pr(B|A)\,Pr(A)}{Pr(B)}$$

(2) applied it to classifn:

$$Pr(Y=k|X=x) = \frac{f_k(x)\,\pi_k}{\sum_{l=1}^{K} f_l(x)\,\pi_l}$$

(3) assume that if $Y=k$, then

$$X \sim N_p(\mu_k, \Sigma).$$

(4) we est. $\hat{\pi}_k, \hat{\mu}_k, \hat{\Sigma}$

(5) we got $\hat{Pr}(Y=k|x)$, and a linear decision boundary

NAIVE BAYES

(1) recalling Bayes:

$$Pr(A|B) = \frac{Pr(B|A)\,Pr(A)}{Pr(B)}$$

(2) applied it to classifn:

$$Pr(Y=k|X=x) = \frac{f_k(x)\,\pi_k}{\sum_{l=1}^{K} f_l(x)\,\pi_l}$$

(3) assume that

$$f_k(x) = \prod_{j=1}^{p} f_{kj}(x_j)$$

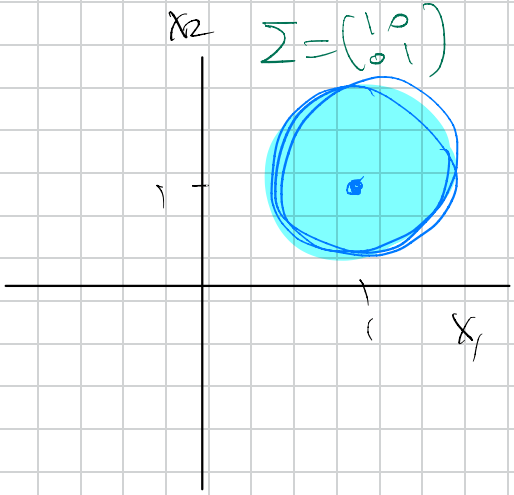
each of the p features is independent

(4) we est. $\{\hat{f}_{kj}(\cdot)\}_{j=1,\dots,p;\,k=1,\dots,K}$

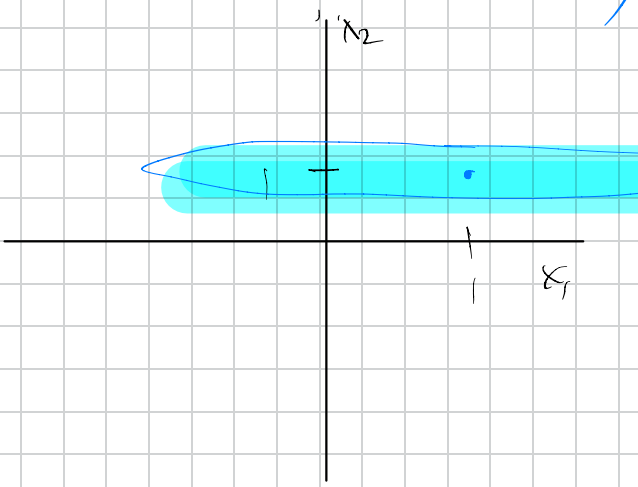
(5) we got $\hat{Pr}(Y=k|x)$

$X \sim N_2(\mu, \Sigma)$; $\mu = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

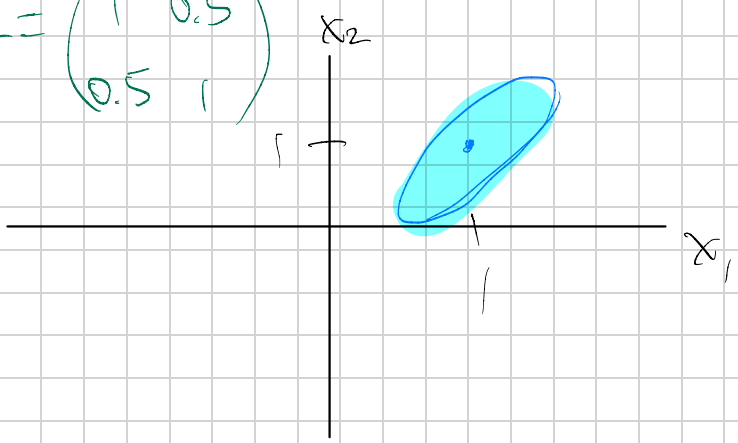
$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$



$\Sigma = \begin{pmatrix} 4 & 0 \\ 0 & 0.2 \end{pmatrix}$



$\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$



Wrapping up Classification

3 options for classification:

1. logistic regression;

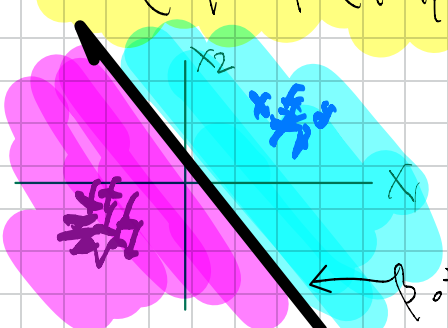
$$\Pr(Y=1|X) = \frac{\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)} \quad (\star)$$

[OR more generally

$$\Pr(Y=1|X) = \frac{\exp(f(X))}{1+\exp(f(X))} \quad (\star\star)$$

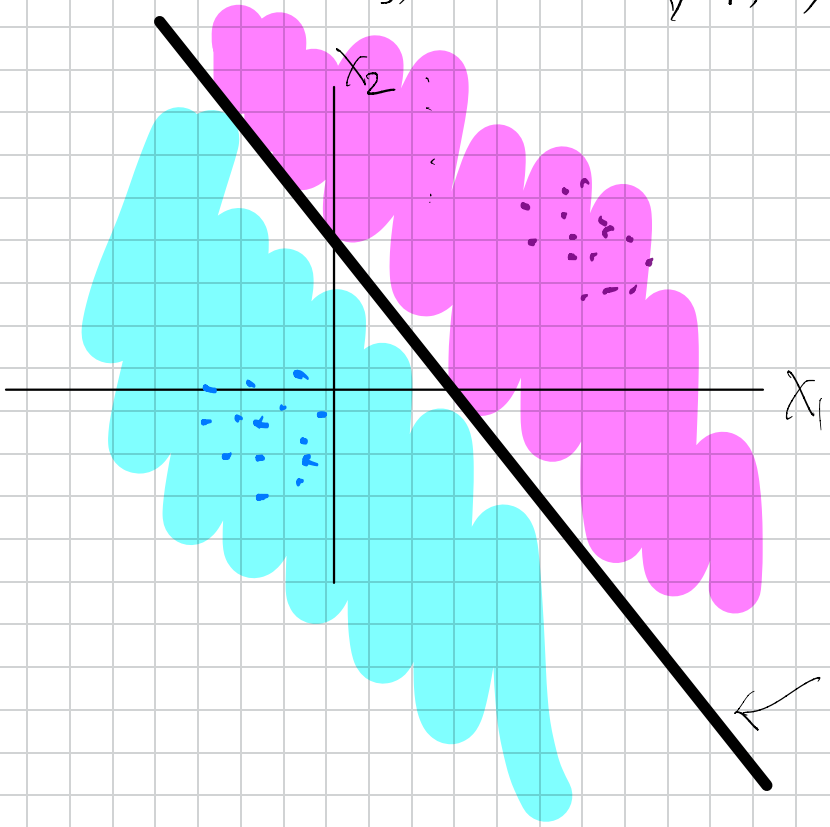
]

$$\log\left(\frac{\Pr(Y=1|X)}{1-\Pr(Y=1|X)}\right) = \begin{cases} \beta_0+\beta_1 X & (\star) \\ f(X) & (\star\star) \end{cases}$$



② linear disc. analysis;

for k^{th} class, $X \sim N(\mu_k, \Sigma)$ } different assumption gives naive Baye



← linear decision boundary

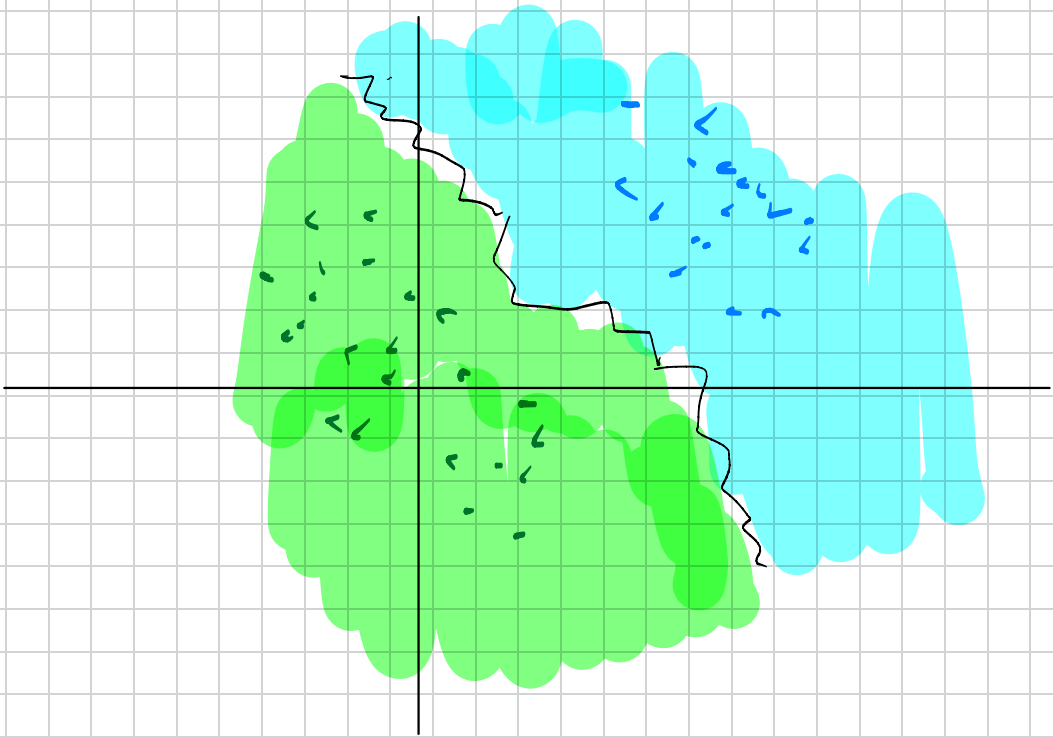
Use it to est

$$Pr(Y=k|X=x)$$

③ K-nearest neighbors

$$\hat{Pr}(Y=l \mid X=x) =$$

among the K nearest neighbors of x, what fraction are in l^{th} class?



Reminder: Bayes classifier is the unattainable ideal — it assigns the obs x to the class k for which

$$\Pr(Y=k \mid X=x) = \frac{\pi_k f_k(x)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(x)}$$

is largest

THE POINT IS: all of these classifiers are doing their best to approx. the Bayes classifier!