CLASSIFICATI

EXAMPLES

- Binary Classification
  - Is email spam or not?
  - Is credit card transaction fraudulent?
  - Is a user male or female?
- Multi-level classification
  - Safety standard of a car.
  - Activity is associated with acceleration
  - What kind of flower is shown in a pictur

THE CLASSIFICATION

- Given
  - $X = (X_1, \ldots, X_p)^T$ random *input* variables
  - Output variable G, taking values in set $\mathcal{G}$ with $1, 2, \ldots, K$.
- Task
  - Given a training set $(g_i, x_i), i = 1, \ldots, N$
  - Find a good approximation for $G(x) = \mathrm{E}(G$
  - Usually among the lines of maximizing $\sum_i \log$ where θ are model parameters.

DECISION BOUND

Decision boundaries are *hypersur*

$$\{x \,|\Pr(G = k|X = x) = \Pr(G =$$

LINEAR CLASSIFI

The classification problem is linear if the decision
two classes k, l,

$$\{x \,|\Pr(G = k|X = x) = \Pr(G =$$

are *linear* in x.

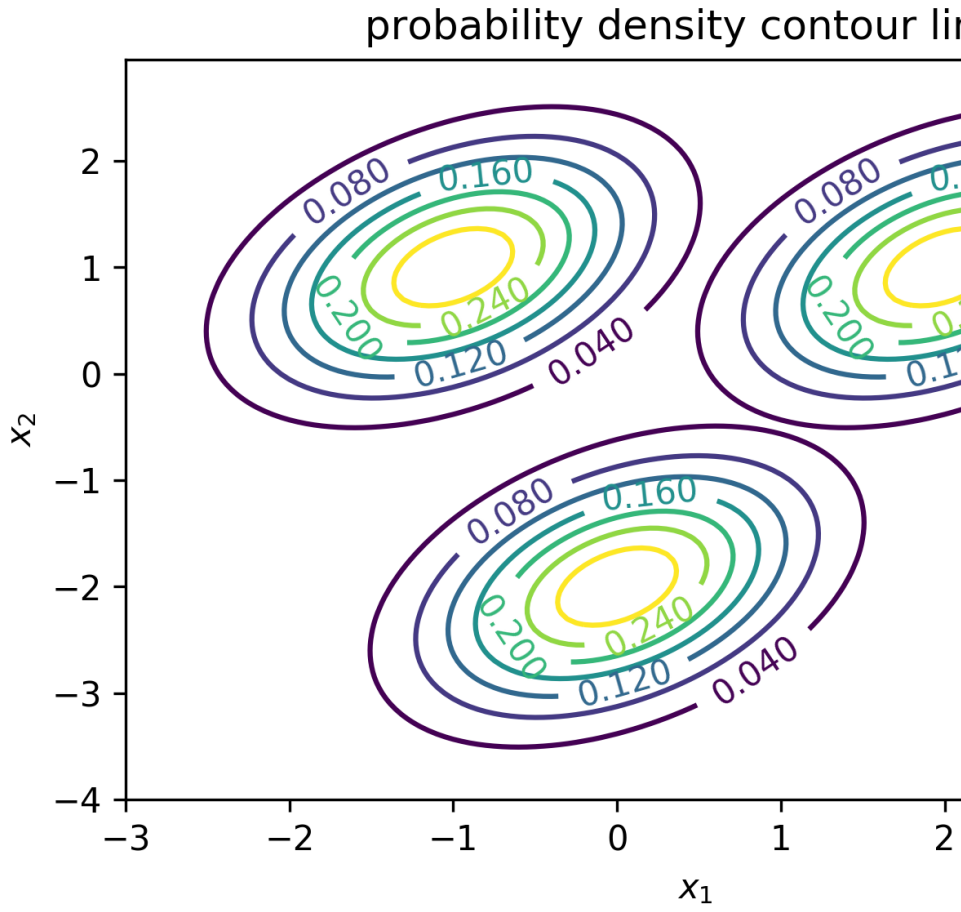
DISCRIMINANT FU

- Popular approach.
- Define *discriminant functions* $\delta_k(x)$
- Classify to $G(x) = \text{argmax}_k \, \delta_k(x$
  - Could e.g. model $\Pr(G = k|X =$
- Decision boundaries are *linear* in x
- The same holds true for $\Pr(G = k$

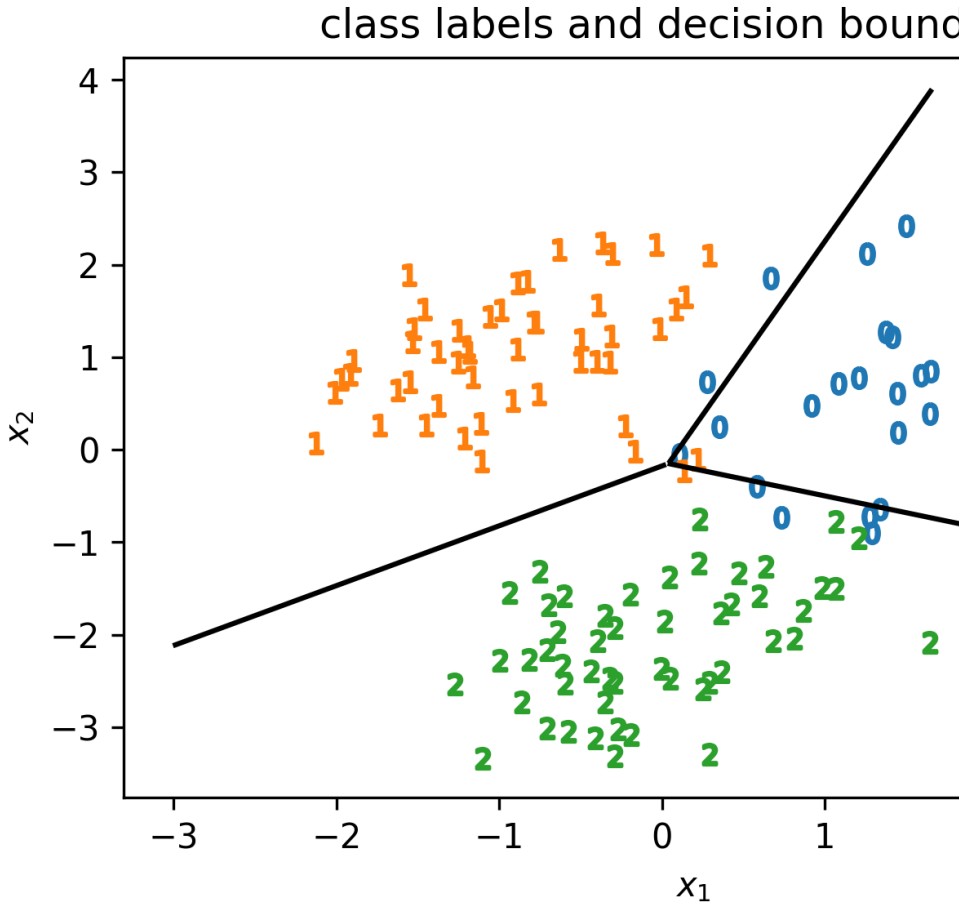
EXAMPLE: MULTIV
NORMAL

$$f(x) = \frac{\exp\left(-\frac{1}{2}(x-\mu_l)^T\Sigma^{-1}\right.}{\sqrt{(2\pi)^d|\Sigma|}}$$

DENSITIES


probability density contour li
0.080
0.160
0.200
0.240
0.120
0.040
2
1
0
−1
−2
−3
−4
−3
−2
−1
0
1
2
x_1
x_2

LINEAR BOUNDA



ACTUALLY ...

... it's enough to have

$$\{x \mid f(\Pr[G = k|X = x]) = f(\Pr[G$$

linear in x / a hyperplane / affine space for

EXAMPLE

$$\log\left(\frac{\Pr[G = k|X = x]}{\Pr[G = l|X = x]}\right) = \theta$$

VARIABLES

- Of course we can still have $X_i = X_j^2$ or X
- Decision boundaries can still be seen as li

CATEGORICAL INP

- *Naïve approach*: Convert to numeric.
  - Generally bad idea.
  - Defining a metric on categoricals tricky.
- *Usual approach*: One-hot encoding.
  - For a K-level categorical, introduce $K-1$

$$x_i = \begin{cases} 1 & \text{if } g_i = \\ 0 & \text{else} \end{cases}$$

  - Effect of k-th variable is the effect of having

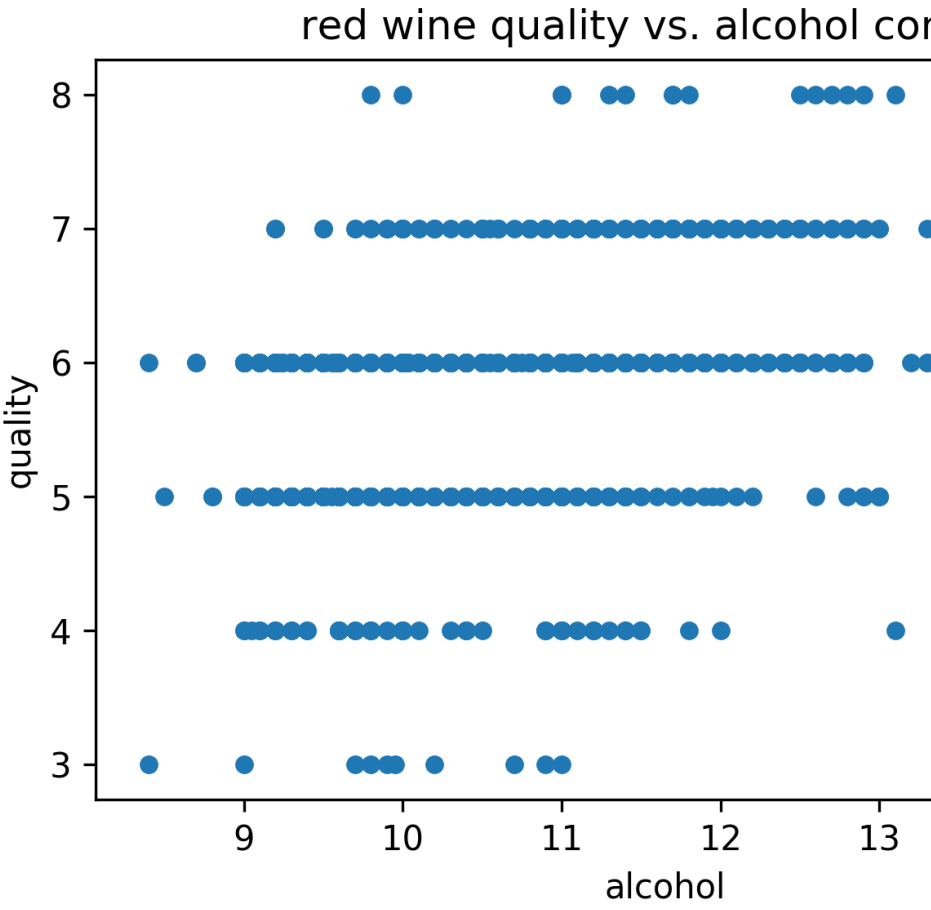
TRAINING DA

- Need to make sure all *classes* are re
- What if one (or more) classes are ov
  - Different priors?
  - Data collection artifact?
  - Re-balance training data?

ORDERED TARG

- What if we have ordered cate
- Sometimes a bit of a moving

RED WINE QUA


red wine quality vs. alcohol co
quality
8
7
6
5
4
3
9
10
11
12
13
alcohol

CATEGORICAL OI

- *Pros*
  - Don't have to think (too much) about metric.
  - Don't have to think (too much) about subject
- *Cons*
  - Using e.g. linear regression, you can answer
    - How many quality points per additional pe
  - Information about order is lost.
  - Sometimes gives less variance.

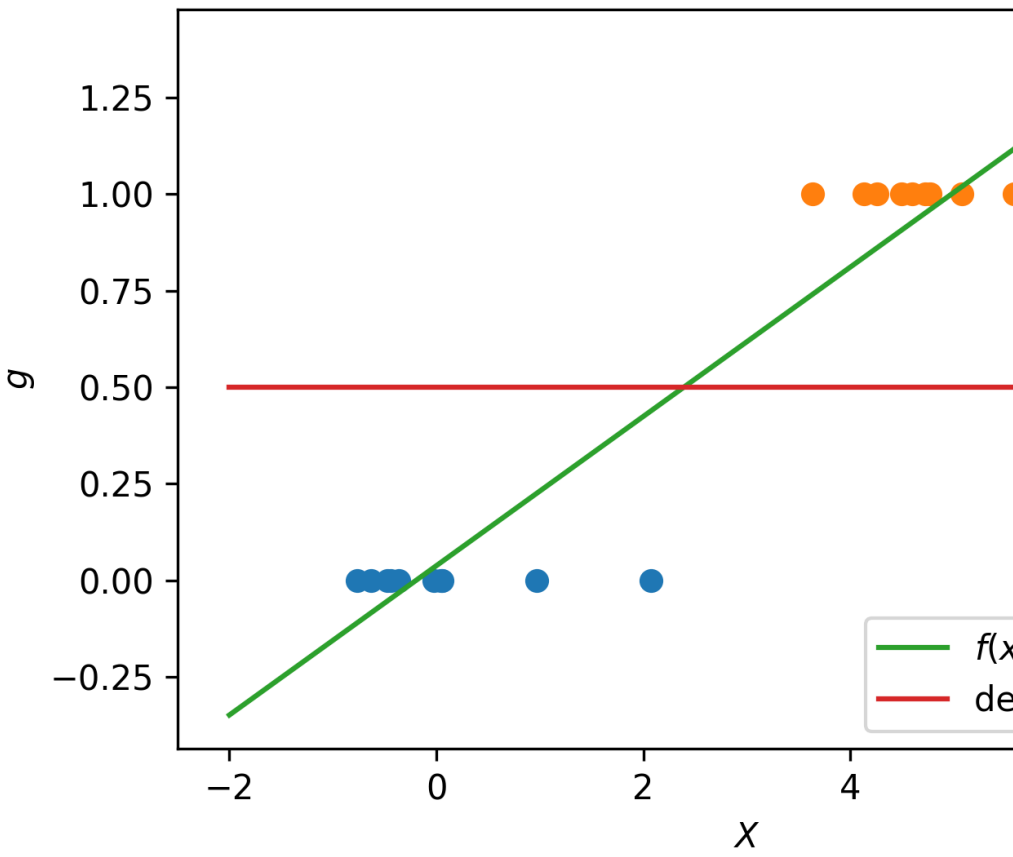
METHODS

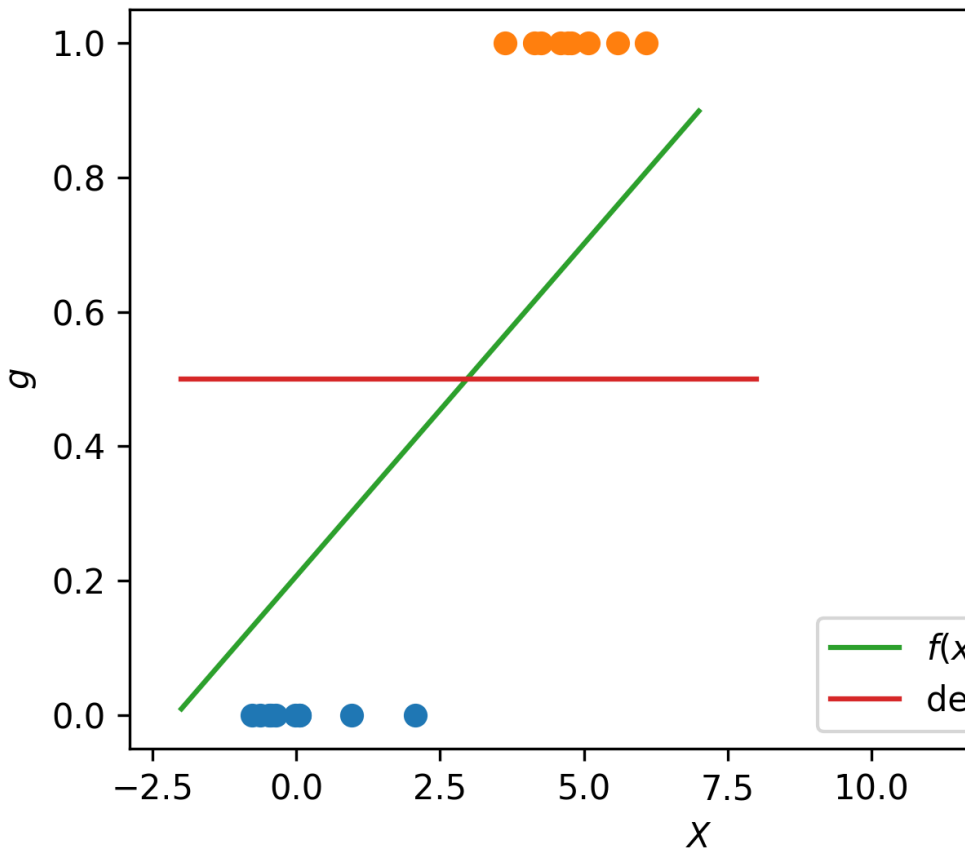
LINEAR REGRES

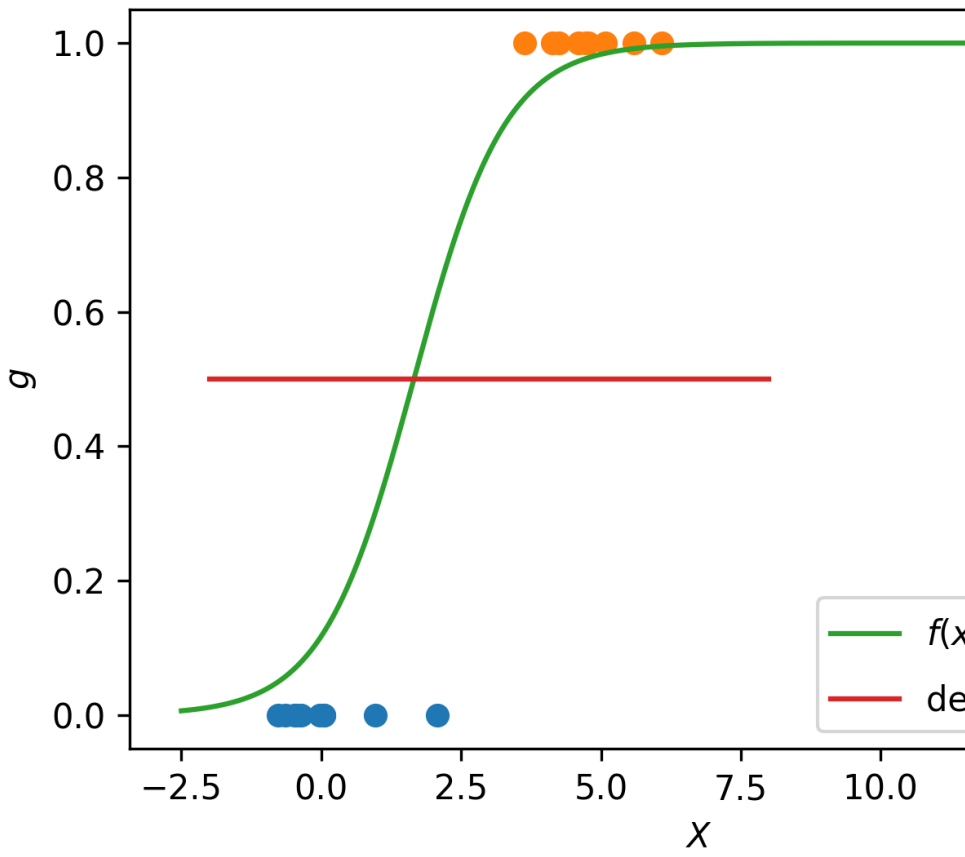
- Write an indicator variable $Y = ($
  - $Y_l = \begin{cases} 1 & \text{if } G = l\,, \\ 0 & \text{else.} \end{cases}$
- Fit a linear model

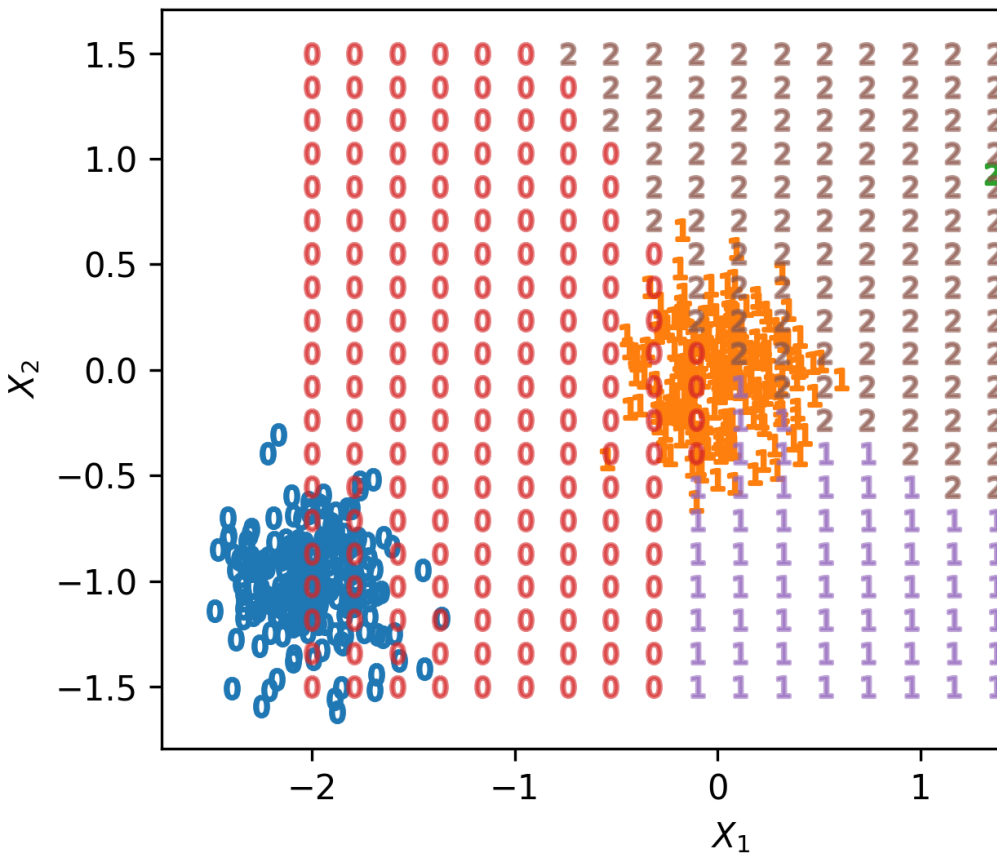
$$\hat{f}_l = \hat{\theta}_{0,l} + \hat{\theta}_l x^T.$$

  to each Y_l
- Classify $\hat{G}(x) = \text{argmax}_{k \in \mathcal{G}}\, \hat{f}_k$
- Can have disastrous results.


1.25
1.00
0.75
0.50
0.25
0.00
−0.25
g
−2
0
2
4
X
f(x
de


1.0
0.8
0.6
0.4
0.2
0.0
g
−2.5
0.0
2.5
5.0
7.5
10.0
X
f(x
de


1.0
0.8
0.6
0.4
0.2
0.0
g
−2.5
0.0
2.5
5.0
7.5
10.0
X
f(x
de


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
X_2
−2
−1
0
1
X_1

LINEAR DISCRIMINAN

Use Bayes' theorem

$$\Pr(G = l|X = x) = \frac{\Pr(X = x|G =}{\Pr(X}$$

$$= \frac{f_l(x)\pi_l}{\sum_m f_m(x)\pi_m}$$

HOW TO MODEL TH
DISTRIBUTIO

The choice of f_l determines whether LD

LDA

$$f_l(x) \propto \exp\left(-\frac{1}{2}(x-\mu_l)^T\Sigma^{-1}\right.$$

QDA

$$f_l(x) \propto \exp\left(-\frac{1}{2}(x-\mu_l)^T\Sigma_l^{-1}\right.$$

COMPUTING L

$$\hat{\pi}_l = \frac{N_l}{N}$$

$$\hat{\mu}_l = \frac{1}{N_l} \sum_{i, g_i = l} x_i$$

$$\hat{\Sigma} = \frac{1}{N - K} \sum_{l} \sum_{i, g_i = l} (x_i - \hat{\mu}_l)$$

$$\hat{\Sigma}_l = \frac{1}{N_l - 1} \sum_{i, g_i = l} (x_i - \hat{\mu}_l)(x_i -$$

DISCRIMINANT FUNCTIO

One can now use

$$\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu$$

or

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-}$$

to classify

$$G(x) = \text{argmax}_k\ \delta_k(x$$

SOME WORDS O

- Using Σ_l does **not** yield linear decision bound
- What if we have one class l, such that $X_i = 0$
- Won't be able to compute Σ_l^{-1}!
- Has many more parameters
  - LDA $(K-1)(p+1)$
  - QDA $(K-1)(p(p+3)//2+1)$

REGULARIZED DISC
ANALYSIS

In some cases (one example: incomplete rank of
use

$$\hat{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}, \quad 0$$

The regularization

$$\hat{\Sigma}(\gamma) = \gamma\hat{\Sigma} + (1-\gamma)\sigma$$

is also sometimes used

WHY SHOULD YOU

PROS

- Simple.
- Fast.
- Powerful.
- Stable.

CONS

- No confidence intervals.
- More work to get predictor im

LOGISTIC REGRE

Model posteriors via linear funct

$$\log\left(\frac{\Pr[G=1|X=x]}{\Pr[G=K|X=x]}\right) = \theta_1$$

$$\log\left(\frac{\Pr[G=2|X=x]}{\Pr[G=K|X=x]}\right) = \theta_2$$

$$\cdots$$

$$\log\left(\frac{\Pr[G=K-1|X=x]}{\Pr[G=K|X=x]}\right) = \theta_($$

LOGISTIC REGRE

This gives us

$$\Pr(G = k|X = x) = \frac{\exp\left(\theta_{k0} + \theta_k^T x\right)}{1 + \sum_{l=1}^{K-1} \exp\left(\theta_{l0} + \theta_l^T\right.}$$

$$\Pr(G = K|X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp\left(\theta_{l0} + \theta_l^T\right.}$$

HOW TO EXTRACT THE PA

Maximum likelihood estimatio

$$l(\theta) = \sum_{i=1}^{N} \log \Pr(G = g_i | X =$$

Find θ using

$$\theta = \underset{\theta}{\operatorname{argmax}}\, l(\theta)$$

via e.g. the Newton-Rhapson

$$\theta_{\text{new}} = \theta_{\text{old}} - \left(\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^T} \right)^{-}$$

EVALUATION OF B
CLASSIFIER

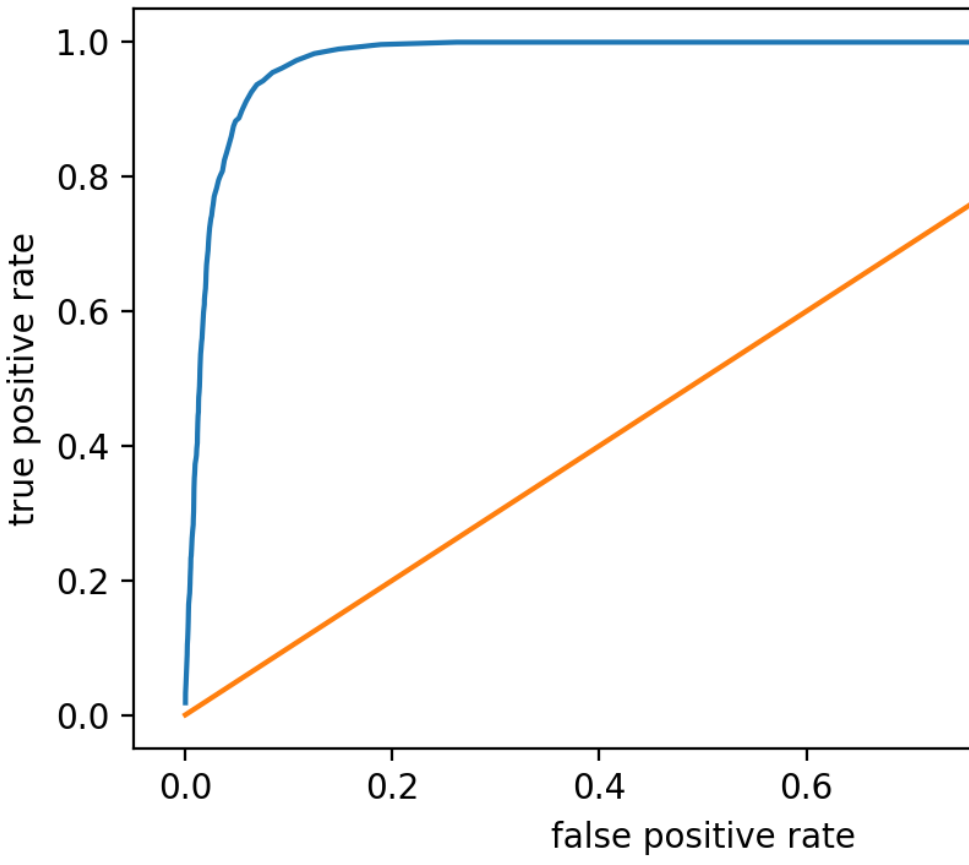
- True positive (TP)
  - Predicted 1, actually
- True negative (TN)
  - Predicted 0, actually
- False positive (FP)
  - Predicted 1, actually
- False negative (FN)
  - Predicted 0, actually

OBJECTIVE

- Sensitivity (true positive rate, hit rate
  - $TPR = TP/P = TP/(TP +$
  - Want to optimize e.g. for tests for
  - Similarly: $FPR = FP/N = FP$
- Specicifity (true negative rate)
  - $TNR = TN/N = TN/(TN +$
  - Want to optimize e.g. in credit ris
- Precision (positive predictive value)
  - $PPV = TP/(TP + FP)$
  - Want to optimize this e.g. for cred

ROC

The receiver operating characteristic plot



AUC

The area under the (ROC) curve (AUC) is a scalar
model quality (for some value o