CLASSIFICATION

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 data from smartphone.
  - What kind of flower is shown in a picture?

THE CLASSIFICATION PROBLEM

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

DECISION BOUNDARIES

Decision boundaries are *hypersurfaces* where

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

LINEAR CLASSIFICATION

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

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

are *linear* in x.

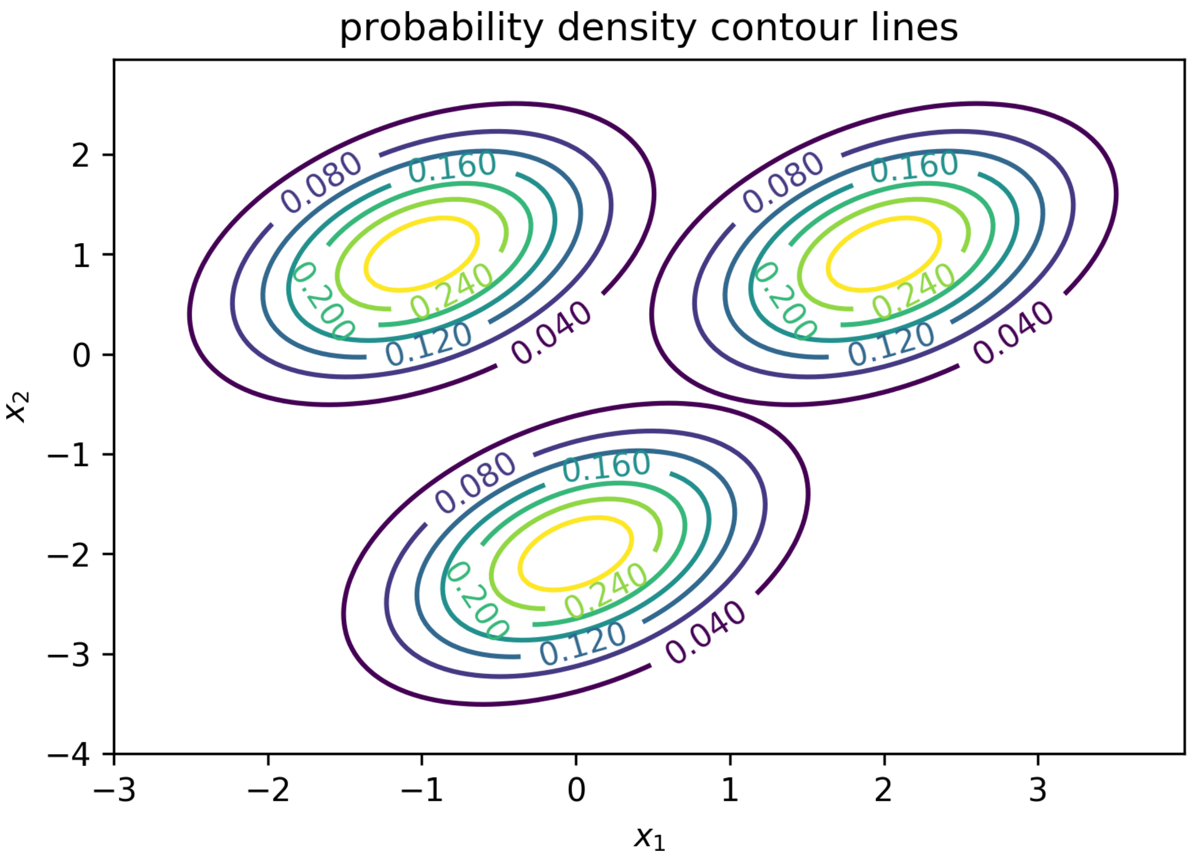
DISCRIMINANT FUNCTIONS

- Popular approach.
- Define *discriminant functions* $\delta_k(x)\, k = 1, \ldots, K$.
- Classify to $G(x) = \mathrm{argmax}_k\, \delta_k(x)$.
    - Could e.g. model $\Pr(G = k|X = x)$ directly.
- Decision boundaries are *linear* in x if $\delta_k(x)$ are.
- The same holds true for $\Pr(G = k|X = x)$.

EXAMPLE: MULTIVARIATE NORMAL

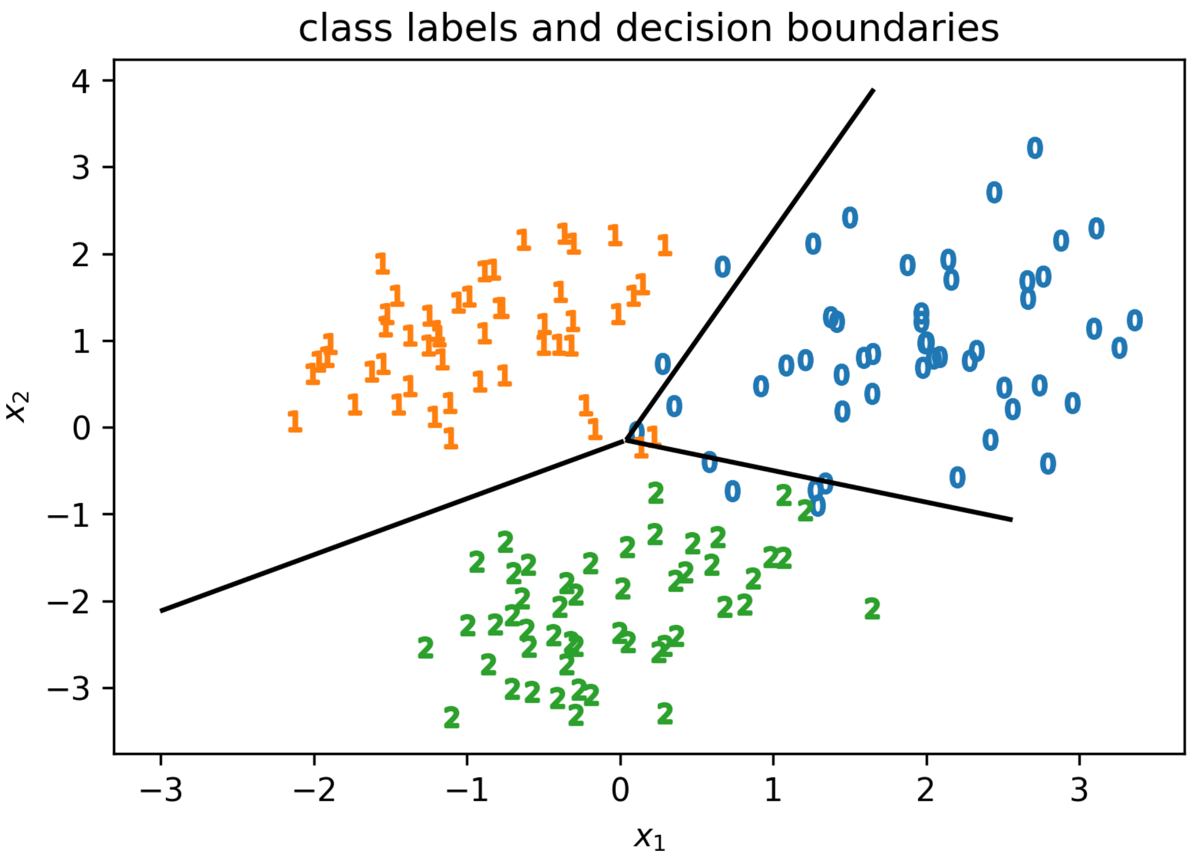
$$f(x) = \frac{\exp\left(-\frac{1}{2}(x-\mu_l)^T\Sigma^{-1}(x-\mu_l)\right)}{\sqrt{(2\pi)^d|\Sigma|}}$$

DENSITIES



probability density contour lines
0.080
0.160
0.200
0.240
0.120
0.040
0.080
0.160
0.200
0.240
0.120
0.040
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
3
x_1
x_2

LINEAR BOUNDARIES



ACTUALLY ...

... it's enough to have

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

linear in x / a hyperplane / affine space for some *monotone* f.

EXAMPLE

$$\log\left(\frac{\Pr[G = k|X = x]}{\Pr[G = l|X = x]}\right) = \theta_0 + \theta^T x$$

VARIABLES

- Of course we can still have $X_i = X_j^2$ or $X_i = X_k\, X_l$.
- Decision boundaries can still be seen as linear in new variables.

CATEGORICAL INPUT DATA

- *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 $K-1$ new variables $X_1, \ldots, X_k$,

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

  - Effect of k-th variable is the effect of having $g_i = k$ instead of $g_i = K$.

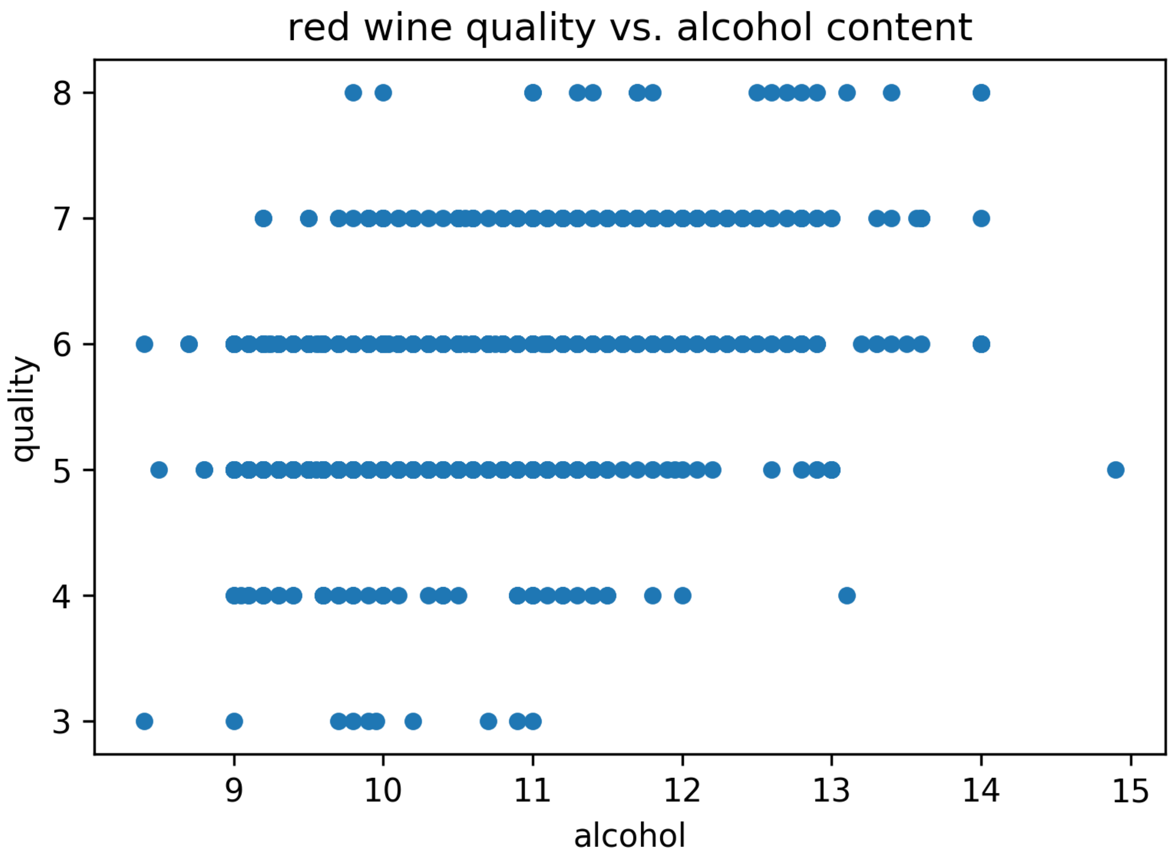
TRAINING DATA

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

ORDERED TARGETS

- What if we have ordered categoricals?
- Sometimes a bit of a moving target ...

RED WINE QUALITY



CATEGORICAL OR NOT?

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

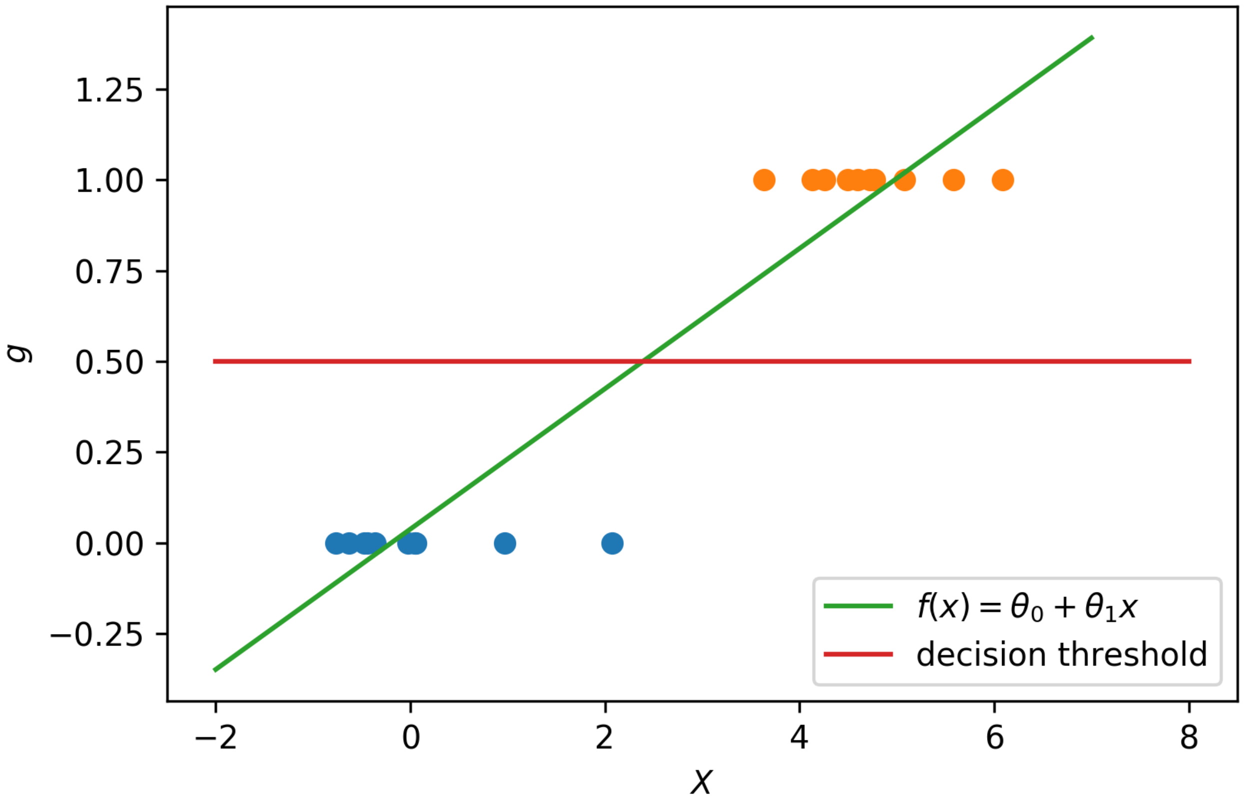
METHODS

LINEAR REGRESSION

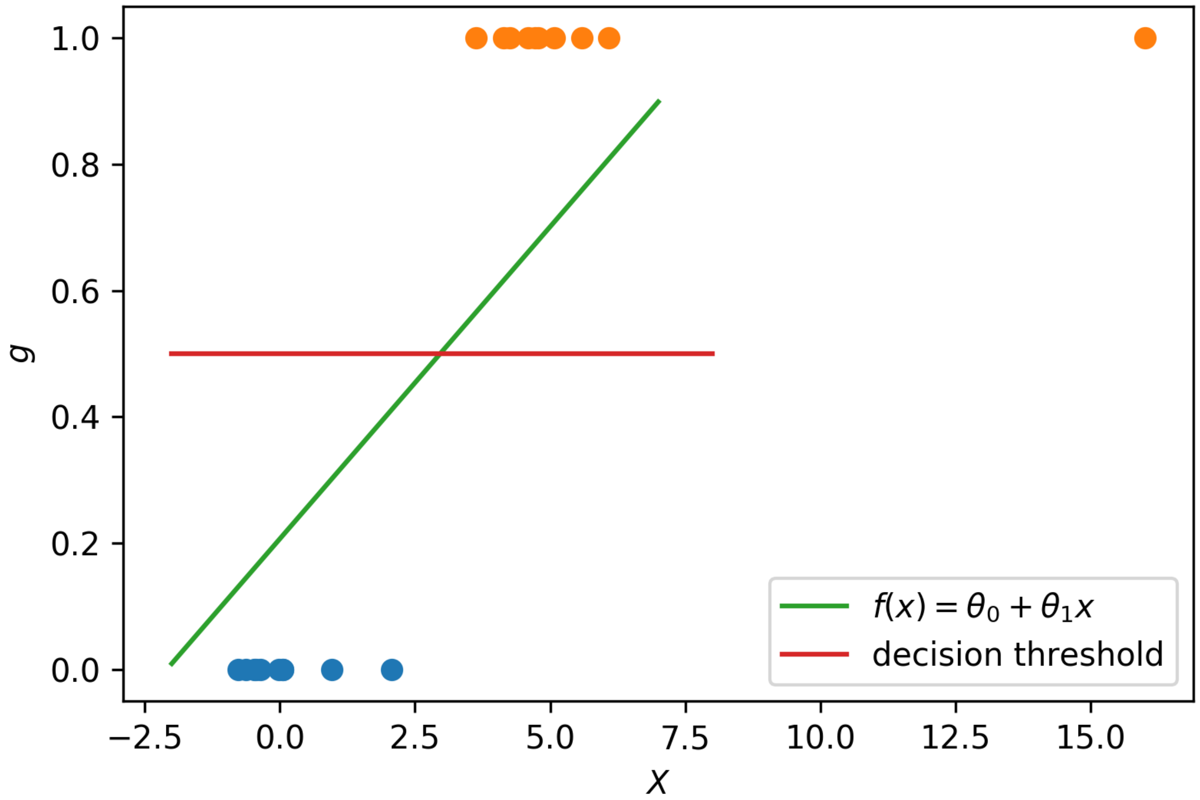
- Write an indicator variable $Y = (Y_0, \ldots, Y_K)$.
    - $Y_l = \begin{cases} 1 & \text{if } G = l\,, \\ 0 & \text{else.} \end{cases}$
- Fit a linear model

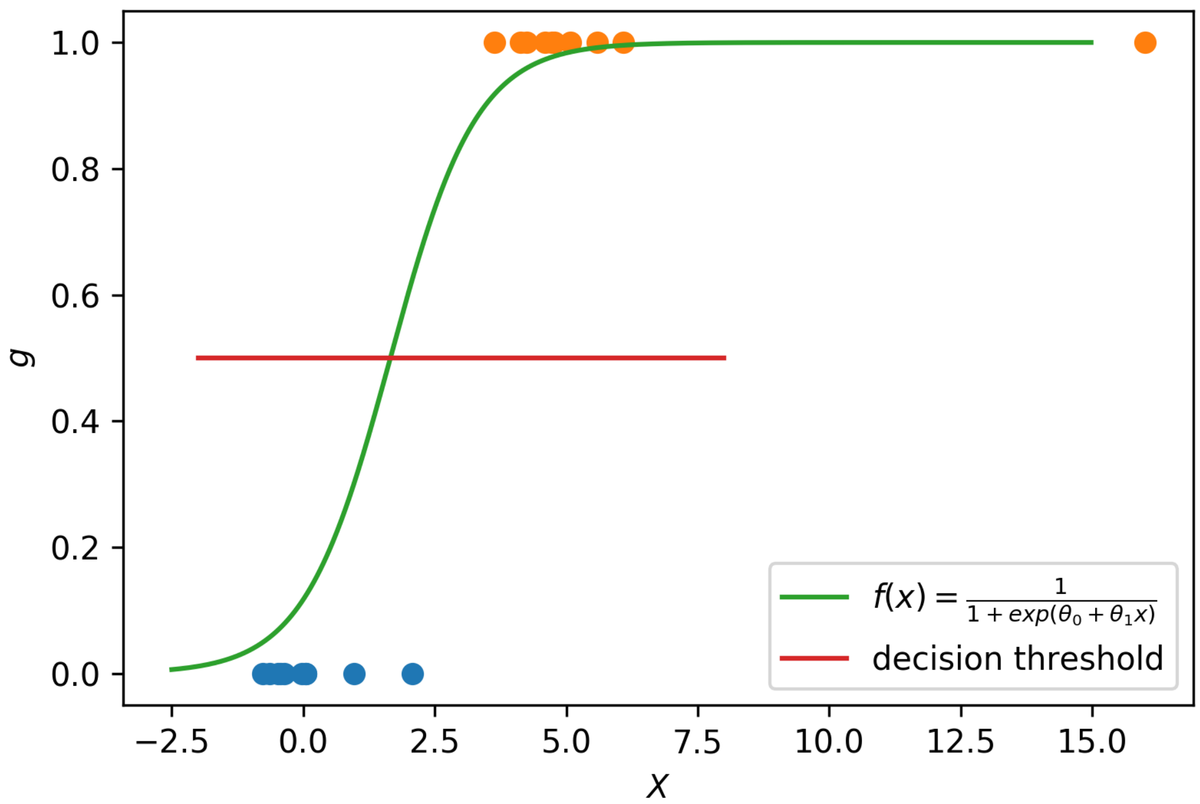
$$\hat{f}_l = \hat{\theta}_{0,l} + \hat{\theta}_l x^T.$$

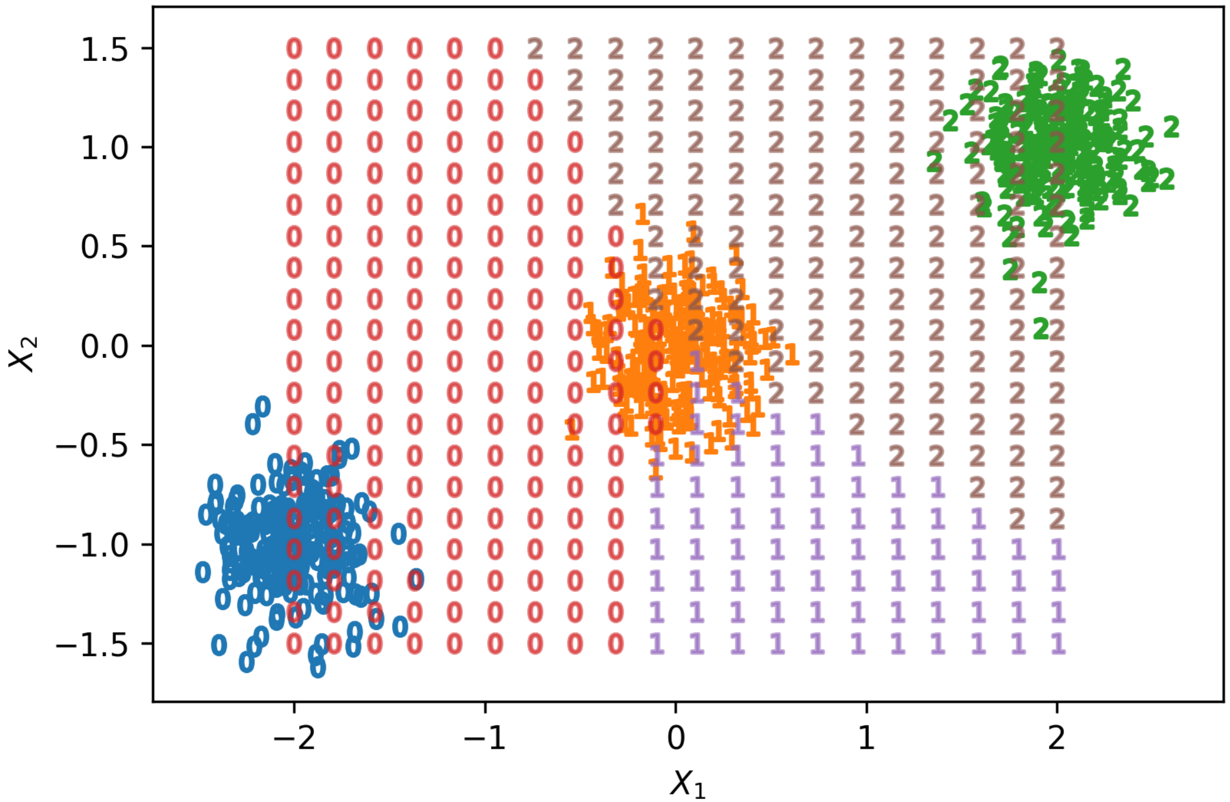
  to each Y_l
- Classify $\hat{G}(x) = \mathrm{argmax}_{k \in \mathcal{G}} \hat{f}_k(x)$.
- Can have disastrous results.



1.25
1.00
0.75
0.50
0.25
0.00
−0.25
g
−2
0
2
4
6
8
X
$f(x) = \theta_0 + \theta_1 x$
decision threshold


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
12.5
15.0
X
$f(x) = \theta_0 + \theta_1 x$
decision threshold


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
12.5
15.0
X
$f(x) = \frac{1}{1+exp(\theta_0+\theta_1 x)}$
decision threshold



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

LINEAR DISCRIMINANT ANALYSIS

Use Bayes' theorem

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

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

HOW TO MODEL THE JOINT DISTRIBUTION?

The choice of f_l determines whether LDA is in fact linear.

LDA

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

QDA

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

COMPUTING LDA

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

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

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

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

DISCRIMINANT FUNCTIONS FOR LDA

One can now use

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

or

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

to classify

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

SOME WORDS ON QDA

- Using Σ_l does **not** yield linear decision boundaries.
- What if we have one class l, such that $X_i = 0$ for all i such that $g_i = l$?
- 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 DISCRIMINANT ANALYSIS

In some cases (one example: incomplete rank of Σ_k), it can be beneficial to use

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

The regularization

$$\hat{\Sigma}(\gamma) = \gamma\hat{\Sigma} + (1-\gamma)\sigma^2 I$$

is also sometimes used.

WHY SHOULD YOU USE LDA?

PROS

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

CONS

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

LOGISTIC REGRESSION

Model posteriors via linear functions in x

$$\log\left(\frac{\Pr[G=1|X=x]}{\Pr[G=K|X=x]}\right) = \theta_{10} + \theta_1^T x$$

$$\log\left(\frac{\Pr[G=2|X=x]}{\Pr[G=K|X=x]}\right) = \theta_{20} + \theta_2^T x$$

$$\cdots$$

$$\log\left(\frac{\Pr[G=K-1|X=x]}{\Pr[G=K|X=x]}\right) = \theta_{(K-1)0} + \theta_{K-1}^T x$$

LOGISTIC REGRESSION

This gives us

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

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

HOW TO EXTRACT THE PARAMETERS?

Maximum likelihood estimation with

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

Find θ using

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

via e.g. the Newton-Rhapson method

$$\theta_{\text{new}} = \theta_{\text{old}} - \left(\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^T} \right)^{-1} \frac{\partial l(\theta)}{\partial\theta}$$

EVALUATION OF BINARY CLASSIFIERS

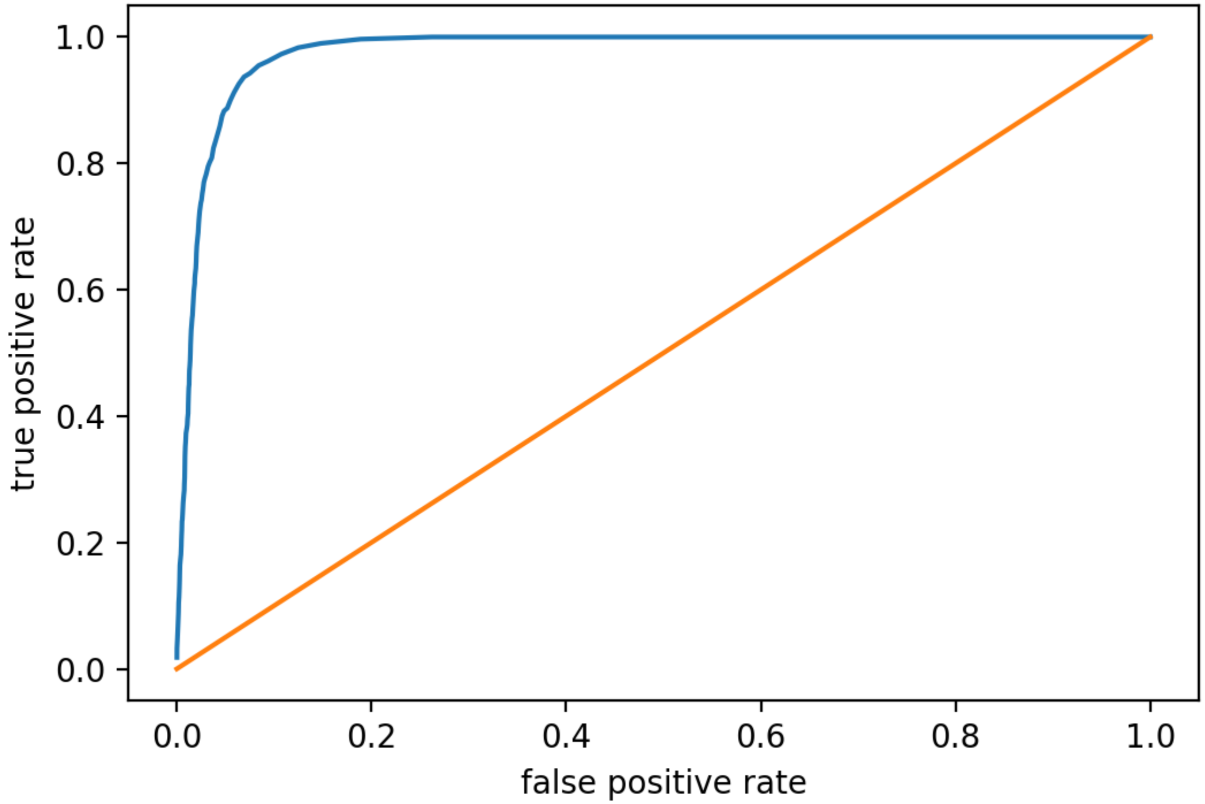
- True positive (TP)
  - Predicted 1, actually 1.
- True negative (TN)
  - Predicted 0, actually 0.
- False positive (FP)
  - Predicted 1, actually 0.
- False negative (FN)
  - Predicted 0, actually 1.

OBJECTIVES

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

ROC

The receiver operating characteristic plots TPR vs. FPR.



AUC

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