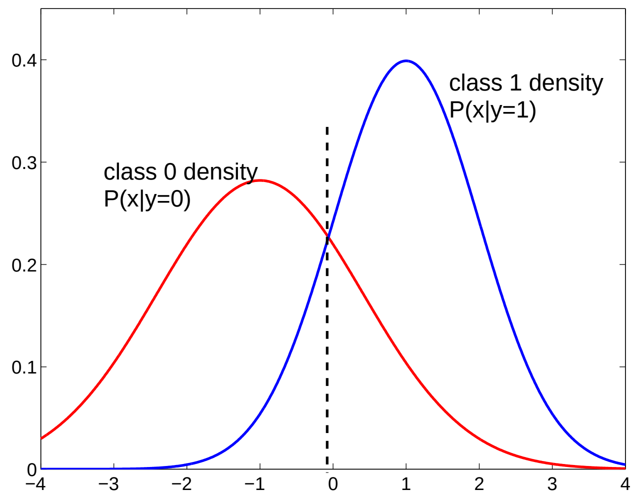
Background: simple decision theory

- Suppose we know the class-conditional densities $p(\mathbf{x}|y)$ for $y = 0, 1$ as well as the overall class frequencies $P(y)$.

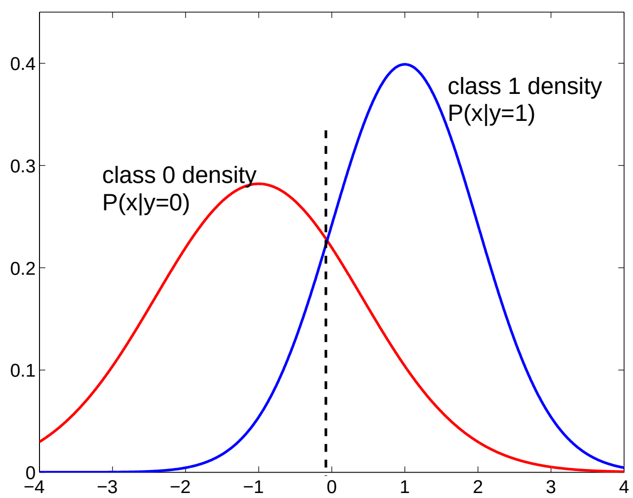
  How do we decide which class a new example $\mathbf{x}'$ belongs to so as to minimize the overall probability of error?



Background: simple decision theory

- Suppose we know the class-conditional densities $p(\mathbf{x}|y)$ for $y = 0, 1$ as well as the overall class frequencies $P(y)$.

  How do we decide which class a new example $\mathbf{x}'$ belongs to so as to minimize the overall probability of error?



The minimum probability of error decisions are given by

$$
\begin{aligned}
y' &= \arg\max_{y=0,1} \{\, p(\mathbf{x}'|y)P(y) \,\} \\
&= \arg\max_{y=0,1} \{\, P(y|\mathbf{x}') \,\}
\end{aligned}
$$

Bayes Classifier (ideal!!)

In general, P(ylx') is not known. Most classifiers provide a MODEL for P(ylx')

Assume for one $(\boldsymbol{x}, y)$ pair:

$$
\begin{aligned}
p(y=1|\boldsymbol{x};\boldsymbol{\theta}) &= h_{\boldsymbol{\theta}}(\boldsymbol{x}) \\
p(y=0|\boldsymbol{x};\boldsymbol{\theta}) &= 1 - h_{\boldsymbol{\theta}}(\boldsymbol{x})
\end{aligned}
$$

More compactly:

$$
p(y|\boldsymbol{x};\boldsymbol{\theta}) = (h_{\boldsymbol{\theta}}(\boldsymbol{x}))^y (1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^{1-y}
$$

then $p(y|\boldsymbol{x};\boldsymbol{\theta}) \sim \text{Bernoulli}(h_{\boldsymbol{\theta}}(\boldsymbol{x}))$

Logistic Regression - Probabilistic interpretation

What decision boundary does this set up lead to?
Another way to think about the prediction is $y = 1$ if

$$\log \frac{p(y=1|\boldsymbol{x};\boldsymbol{\theta})}{p(y=0|\boldsymbol{x};\boldsymbol{\theta})} > 0.$$

and 0 otherwise.
Plugging in our hypothesis leads to a linear decision boundary:

$$\log \frac{h_{\boldsymbol{\theta}}(\boldsymbol{x})}{1 - h_{\boldsymbol{\theta}}(\boldsymbol{x})} = \boldsymbol{\theta}^T \boldsymbol{x}$$

You will show this in your homework assignment!

Logistic Regression - Probabilistic interpretation

- Decision theory for binary classification: we assign $\boldsymbol{x}$ to
- the label 1 if
  $p(y=1|\boldsymbol{x}) > p(y=0|\boldsymbol{x})$ (details later)

- For our binary variables $E[y|\boldsymbol{x}] = p(y=1|\boldsymbol{x};\boldsymbol{\theta}) = h_{\boldsymbol{\theta}}(\boldsymbol{x})$
- So $\begin{cases} 1 & \text{if } h_{\boldsymbol{\theta}}(\boldsymbol{x}) > 1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}) \\ 0 \text{ if } h_{\boldsymbol{\theta}}(\boldsymbol{x}) < 1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}) \end{cases}$
- The decision boundary will be marked by $h_{\boldsymbol{\theta}}(\boldsymbol{x}) = 1/2$.

Logistic Regression - Probabilistic interpretation

Given X (the design matrix, which contains all the $x^{(i)}$'s) and θ, what is the distribution of the $y^{(i)}$'s? The probability of the data is given by $p(y|X;\theta)$. This quantity is typically viewed a function of y (and perhaps X), for a fixed value of θ. When we wish to explicitly view this as a function of θ, we will instead call it the likelihood function:

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= p(\boldsymbol{y}|X;\boldsymbol{\theta}) \\
&= \prod_{i=1}^{m} p(y^{(i)}|\boldsymbol{x}^{(i)};\boldsymbol{\theta}) \\
&= \prod_{i=1}^{m} (h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})^{y^{(i)}} (1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}))^{1-y^{(i)}}
\end{aligned}
$$

Logistic Regression - Probabilistic interpretation

Now take the log likelihood:

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \log \mathcal{L}(\boldsymbol{\theta}) \\ &= \sum_{i=1}^{m} y^{(i)} \log h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) + (1 - y^{(i)}) \log(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}))\end{aligned}$$

The negative of the J(θ) we found previously!
minimize J(θ) = maximize $\ell(\boldsymbol{\theta})$

The log-likelihood function is a jointly concave function of the parameters $\boldsymbol{\theta}$;

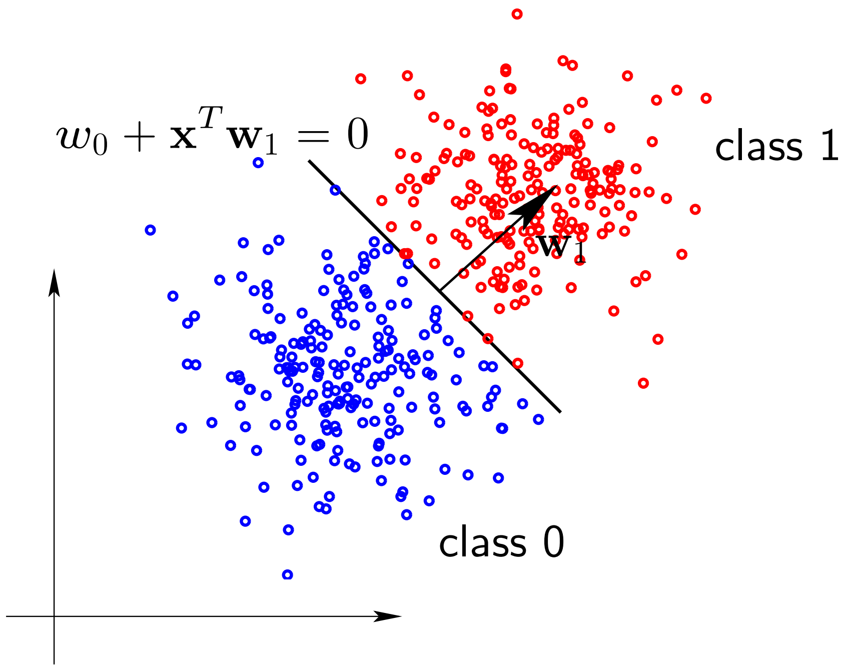
Logistic regression: decisions

- Logistic regression models imply a linear decision boundary

$$\log \frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})} = \boldsymbol{\theta}^T \mathbf{x} \quad = 0 \qquad\qquad \boldsymbol{\theta} \text{ -> } (w_0\,,\, \mathbf{w}_1)$$

$$= w_0 + \mathbf{x}^T \mathbf{w}_1 = 0$$

more interpretable way of representing hyperplanes



Logistic Regression

We have seen that logistic regression finds a linear boundary like so:

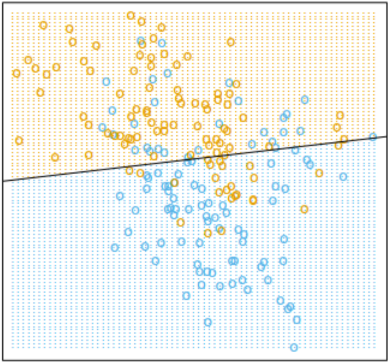


Figure 2: Linear Classifier on data made from 10 bivariate Gaussians with unit variance and different means.