CS 412

APRIL 17TH – MAXIMUM LIKELIHOOD

Maximum Likelihood Learning

Statistical learning

- Want to learn underlying (Gaussian) distributions

Estimating statistics (mean and variance)

Frequentist (MLE) vs. Bayesian (MAP) learning

Inferring distributions

Learning Gaussians from Data

Suppose you have x1, x2, … xR ~ (i.i.d) N(μ,σ2)

But you don't know μ

MLE: For which μ is $x_1, x_2, \ldots x_R$ most likely?

MAP: Which μ maximizes $p(\mu | x_1, x_2, \ldots x_R, \sigma^2)$?

Learning Gaussians from Data

Suppose you have x1, x2, … xR ~ (i.i.d) N(μ,σ2)

But you don't know μ

Maximum Likelihood Estimator: For which μ is $x_1, x_2, \ldots x_R$ most likely?

- Frequentist

Maximum A Posteriori: Which μ maximizes $p(\mu | x_1, x_2, \ldots x_R, \sigma^2)$?

- Bayesian

MLE for univariate Gaussian

Suppose you have x1, x2, … xR ~(i.i.d) N(μ,σ2)

But you don't know μ

MLE: For which μ is x1, x2, … xR most likely?

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2)$$

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2)$$

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2)$$

$$= \arg\max_{\mu} \prod_{i=1}^{R} p(x_i \mid \mu, \sigma^2)$$

(by i.i.d)

$$= \arg\max_{\mu} \sum_{i=1}^{R} \log p(x_i \mid \mu, \sigma^2)$$

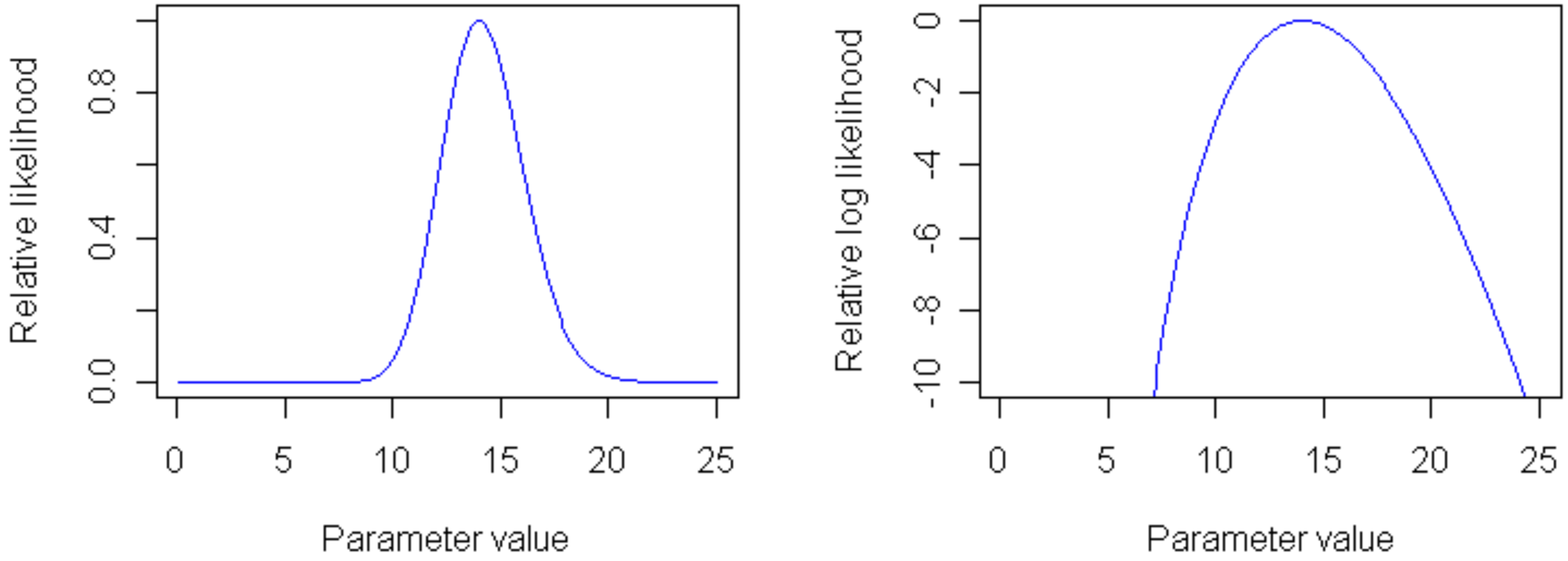
(monotonicity of log)

$$= \arg\max_{\mu} \frac{1}{\sqrt{2\pi}\,\sigma} \sum_{i=1}^{R} - \frac{(x_i - \mu)^2}{2\sigma^2}$$

(plug in formula for Gaussian)

$$\stackrel{\star}{=} \arg\min_{\mu} \sum_{i=1}^{R} (x_i - \mu)^2$$

(after simplification)


Relative likelihood
0.0
0.4
0.8
0
5
10
15
20
25
Parameter value
Relative log likelihood
0
-2
-4
-6
-8
-10
0
5
10
15
20
25
Parameter value

MLE

The best estimate for the mean of the population is the mean of the sample

$$\mu^{mle} = \frac{1}{R}\sum_{i=1}^{R} x_i$$

That matches our expectations!

We've already been using "maximum likelihood" with linear regression

MLE for univariate Gaussian

Suppose you have x1, x2, … xR ~(i.i.d) N(μ,σ2)

But you don't know μ or σ2

MLE: For which q =(μ,σ2) is x1, x2,…xR most likely?

$$\mu^{mle} = \frac{1}{R}\sum_{i=1}^{R} x_i$$

$$\sigma^2_{mle} = \frac{1}{R}\sum_{i=1}^{R} (x_i - \mu^{mle})^2$$

Unbiased Estimators

An estimator of a parameter is unbiased if the expected value of the estimate is the same as the true value of the parameters.

If x1, x2, … xR ~(i.i.d) N(μ,σ2) then

$$E[\mu^{mle}] = E\left[\frac{1}{R}\sum_{i=1}^{R} x_i\right] = \mu$$

μ^{mle} is unbiased

Biased Estimators

An estimator of a parameter is biased if the expected value of the estimate is different from the true value of the parameters.

If x1, x2, … xR ~(i.i.d) N(μ,σ2) then

$$E\left[\sigma^2_{mle}\right] = E\left[\frac{1}{R}\sum_{i=1}^{R}(x_i - \mu^{mle})^2\right] = E\left[\frac{1}{R}\left(\sum_{i=1}^{R}x_i - \frac{1}{R}\sum_{j=1}^{R}x_j\right)^2\right] \neq \sigma^2$$

σ^2_{mle} is biased

MLE Variance Bias

If x1, x2, … xR ~(i.i.d) N(μ,σ2) then

$$E\left[\sigma^2_{mle}\right] = E\left[\frac{1}{R}\left(\sum_{i=1}^{R} x_i - \frac{1}{R}\sum_{j=1}^{R} x_j\right)^2\right] = \left(1 - \frac{1}{R}\right)\sigma^2 \neq \sigma^2$$

Intuition check: consider the case of R=1

Why should our guts expect that σ^2_{mle} would be an underestimate of true σ^2?

Unbiased estimate of Variance

If x1, x2, … xR ~(i.i.d) N(μ,σ2) then

$$E\left[\sigma^2_{mle}\right] = E\left[\frac{1}{R}\left(\sum_{i=1}^{R} x_i - \frac{1}{R}\sum_{j=1}^{R} x_j\right)^2\right] = \left(1 - \frac{1}{R}\right)\sigma^2 \neq \sigma^2$$

So define $\sigma^2_{\text{unbiased}} = \dfrac{\sigma^2_{mle}}{\left(1 - \dfrac{1}{R}\right)}$ So $E\left[\sigma^2_{\text{unbiased}}\right] = \sigma^2$

Unbiased estimate of Variance

If x1, x2, … xR ~(i.i.d) N(μ,σ2) then

$$E\left[\sigma^2_{mle}\right] = E\left[\frac{1}{R}\left(\sum_{i=1}^{R} x_i - \frac{1}{R}\sum_{j=1}^{R} x_j\right)^2\right] = \left(1 - \frac{1}{R}\right)\sigma^2 \neq \sigma^2$$

So define $\sigma^2_{\text{unbiased}} = \dfrac{\sigma^2_{mle}}{\left(1 - \frac{1}{R}\right)}$ So $E\left[\sigma^2_{\text{unbiased}}\right] = \sigma^2$

$$\sigma^2_{\text{unbiased}} = \frac{1}{R-1}\sum_{i=1}^{R}(x_i - \mu^{mle})^2$$

Unbiased estimation

Which is best?

- It depends on the task
- And doesn't make much difference once R--> large

$$\sigma^2_{mle} = \frac{1}{R}\sum_{i=1}^{R}(x_i - \mu^{mle})^2$$

$$\sigma^2_{\text{unbiased}} = \frac{1}{R-1}\sum_{i=1}^{R}(x_i - \mu^{mle})^2$$

Unbiased estimation

- *Assume* $x_1, x_2, \ldots x_R$ ~(i.i.d) $N(\mu,\sigma^2)$
- Suppose we had these estimators for the mean

$$\mu^{suboptimal} = \frac{1}{R+7\sqrt{R}}\sum_{i=1}^{R} x_i$$

$$\mu^{crap} = x_1$$

Are either of these unbiased?

Will either of them asymptote to the correct value as R gets large?

Which is more useful?

MLE for m-dimensional Gaussian

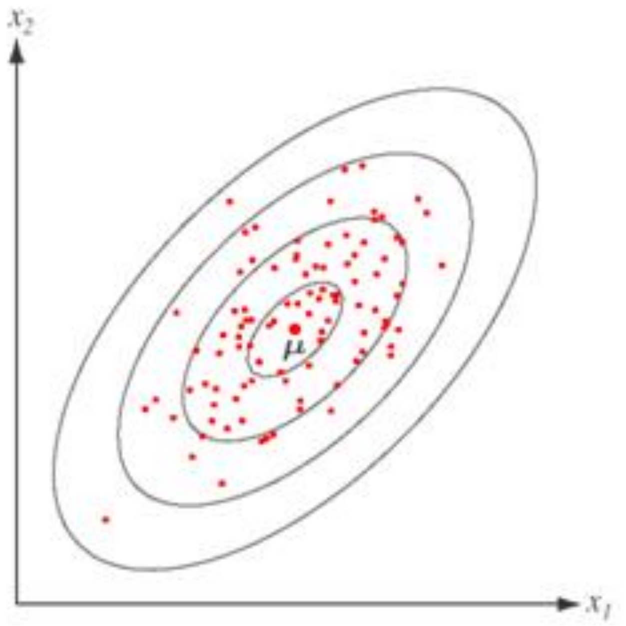
Suppose you have x1, x2, … xR ~(i.i.d) N(μ,S)

But you don't know μ or S

MLE: For which q =(μ,S) is x1, x2, … xR most likely?

$$\boldsymbol{\mu}^{mle} = \frac{1}{R}\sum_{k=1}^{R}\mathbf{x}_k$$

$$\boldsymbol{\Sigma}^{mle} = \frac{1}{R}\sum_{k=1}^{R}\left(\mathbf{x}_k - \boldsymbol{\mu}^{mle}\right)\left(\mathbf{x}_k - \boldsymbol{\mu}^{mle}\right)^T$$



MLE for m-dimensional Gaussian

Suppose you have x1, x2, … xR ~(i.i.d) N(μ,S)

But you don't know μ or S

MLE: For which q =(μ,S) is x1, x2, … xR most likely?

$$\boldsymbol{\mu}^{mle} = \frac{1}{R}\sum_{k=1}^{R}\mathbf{x}_k \qquad \mu_i^{mle} = \frac{1}{R}\sum_{k=1}^{R}\mathbf{x}_{ki}$$

$$\boldsymbol{\Sigma}^{mle} = \frac{1}{R}\sum_{k=1}^{R}\left(\mathbf{x}_k - \mu^{mle}\right)\left(\mathbf{x}_k - \mu^{mle}\right)^T$$

Where $1 \leq i \leq m$

And x_{ki} is value of the i^{th} component of $\mathbf{x}_k$ (the i^{th} attribute of the k^{th} record)

And μ_i^{mle} is the i^{th} component of $\boldsymbol{\mu}^{mle}$

MLE for m-dimensional Gaussian

Suppose you have x1, x2, … xR ~(i.i.d) N(μ,S)

But you don't know μ or S

MLE: For which q =(μ,S) is x1, x2, … xR most likely?

Where $1 \le i \le m$, $1 \le j \le m$

And x_{ki} is value of the i^th^ component of $\mathbf{x}_k$ (the i^th^ attribute of the k^th^ record)

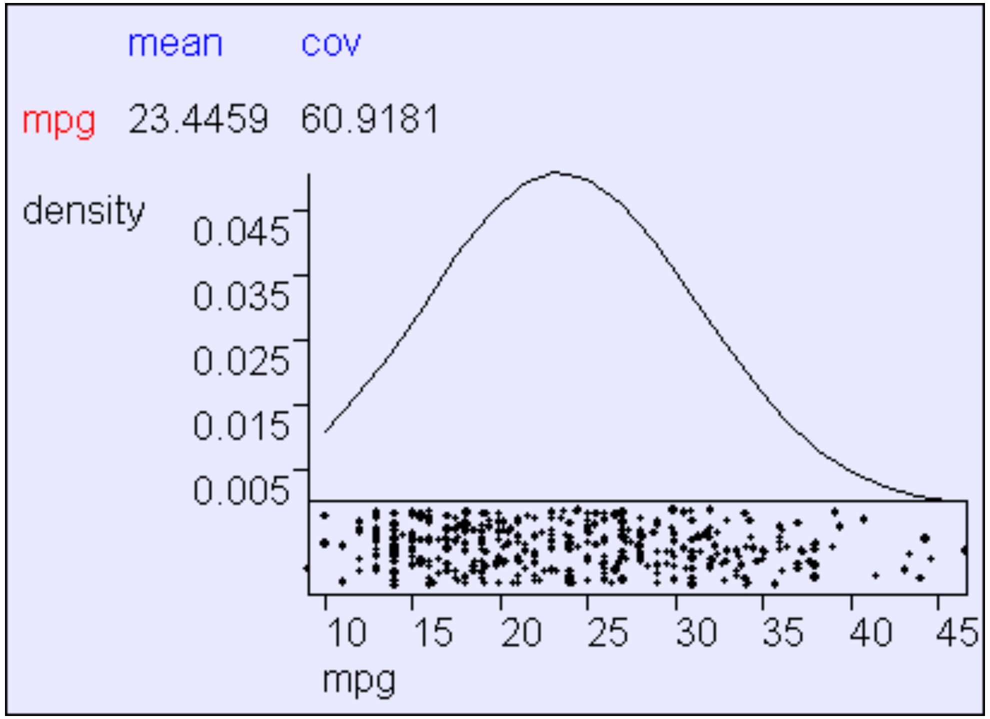
And σ_{ij}^{mle} is the (i,j)^th^ component of $\mathbf{\Sigma}^{mle}$

$$\boldsymbol{\mu}^{mle} = \frac{1}{R}\sum_{k=1}^{R}\mathbf{x}_k \qquad \mu_i^{mle} = \frac{1}{R}\sum_{k=1}^{R}\mathbf{x}_{ki}$$

$$\mathbf{\Sigma}^{mle} = \frac{1}{R}\sum_{k=1}^{R}\left(\mathbf{x}_k - \mu^{mle}\right)\left(\mathbf{x}_k - \mu^{mle}\right)^T$$

$$\sigma_{ij}^{mle} = \frac{1}{R}\sum_{k=1}^{R}\left(\mathbf{x}_{ki} - \mu_i^{mle}\right)\left(\mathbf{x}_{kj} - \mu_j^{mle}\right)$$

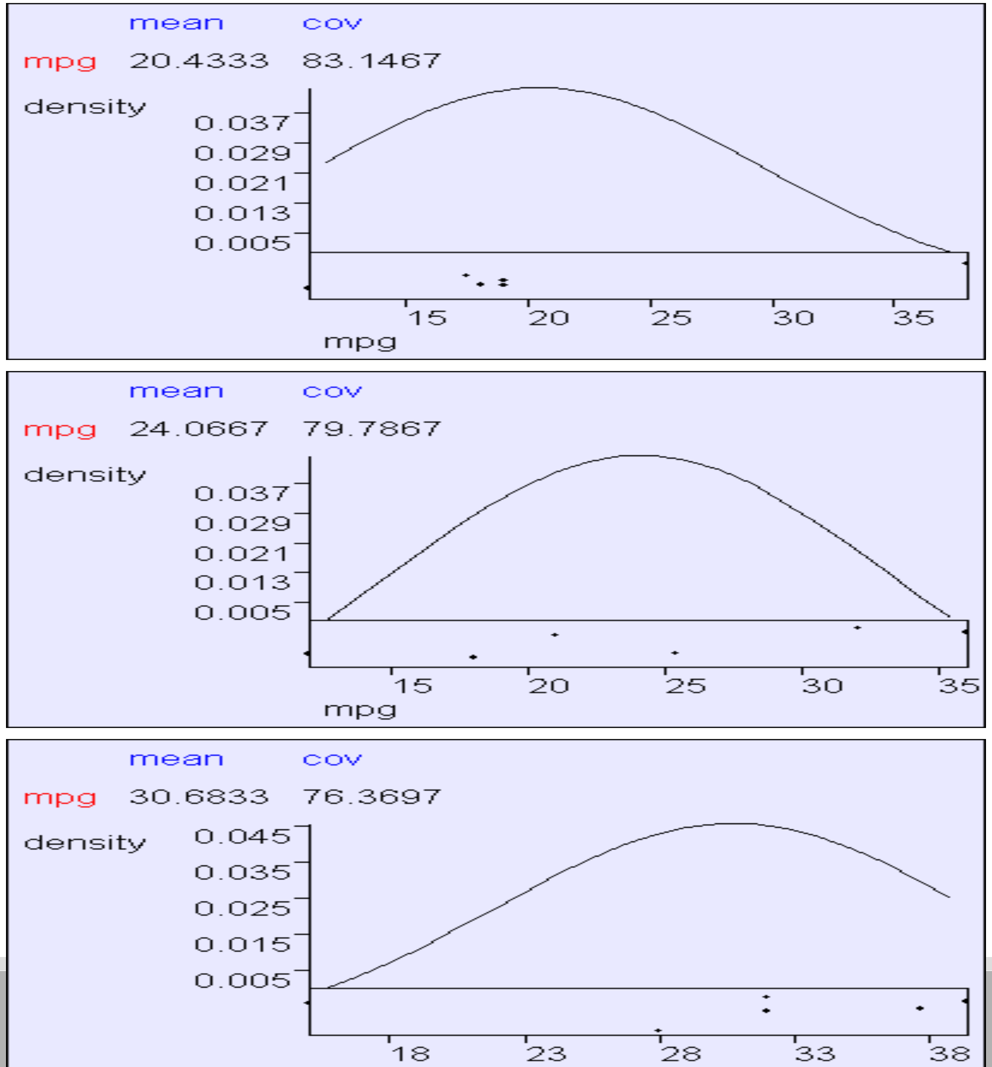
Gaussian MLE in action



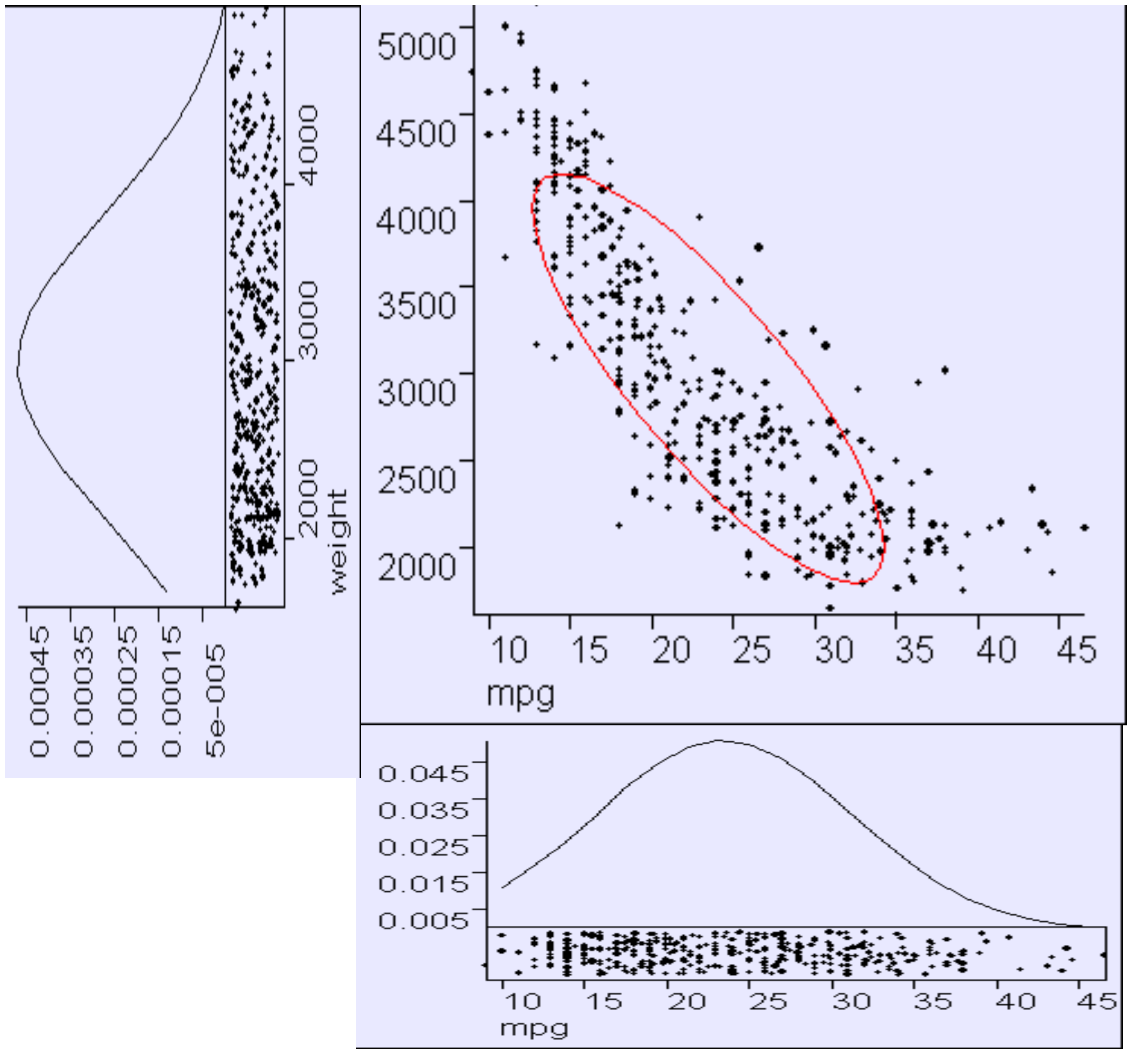
Data-starved Gaussian MLE

Using three subsets of MPG.

Each subset has 6 randomly-chosen cars.



Bivariate MLE in action



Multivariate Data

Multiple measurements (sensors)

d inputs/features/attributes: d-variate

N instances/observations/examples

$$\mathbf{X} = \begin{bmatrix} X_1^1 & X_2^1 & \cdots & X_d^1 \\ X_1^2 & X_2^2 & \cdots & X_d^2 \\ \vdots & & & \\ X_1^N & X_2^N & \cdots & X_d^N \end{bmatrix}$$

Multivariate Parameters

$$\Sigma \equiv \mathrm{Cov}(\boldsymbol{X}) = E\left[(\boldsymbol{X}-\mu)(\boldsymbol{X}-\mu)^T\right] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & & & \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{bmatrix}$$

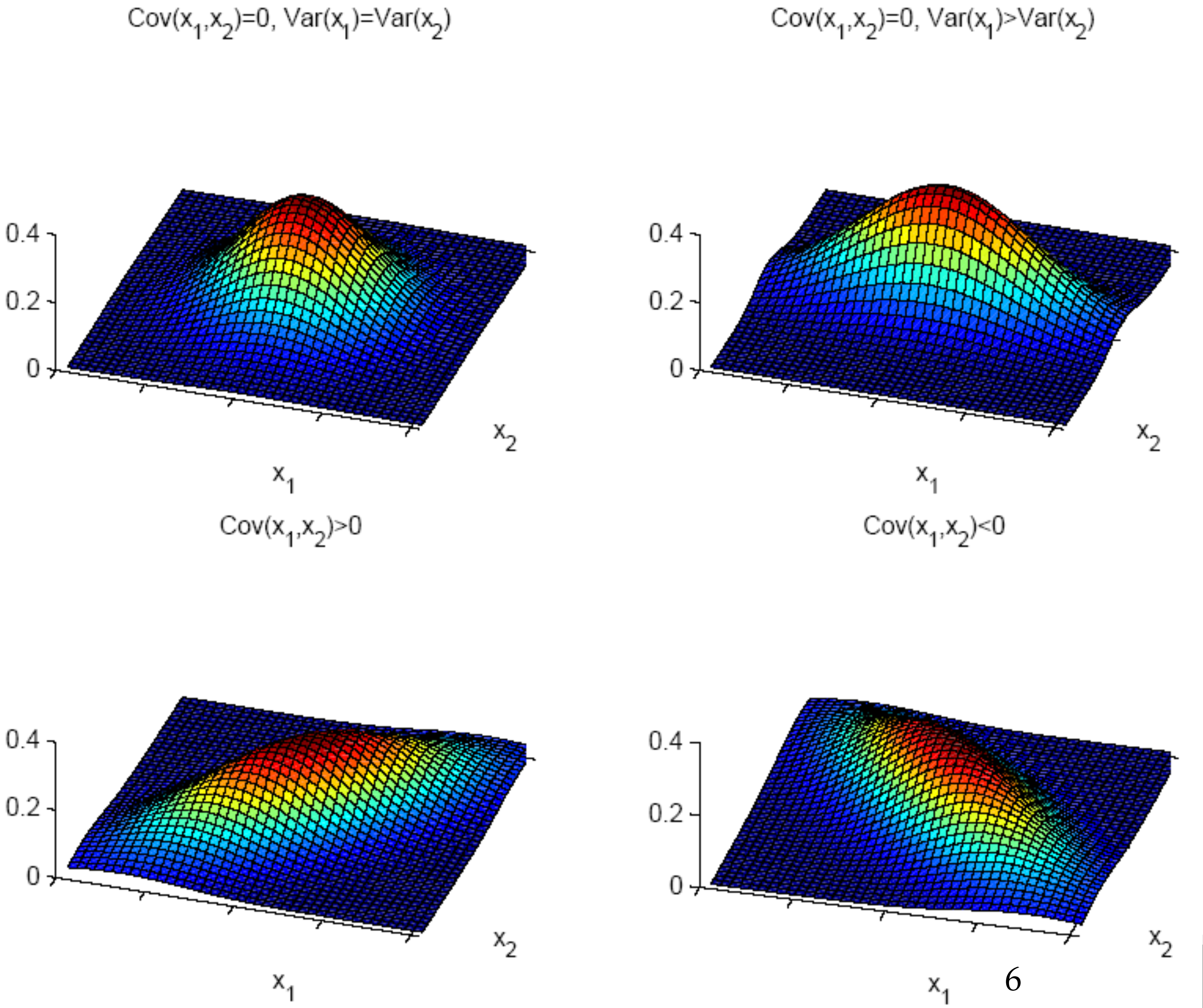
$$\text{Mean}: E[\boldsymbol{x}] = \mu = [\mu_1, \ldots, \mu_d]^T$$

$$\text{Covariance}: \sigma_{ij} \equiv \mathrm{Cov}(X_i, X_j)$$

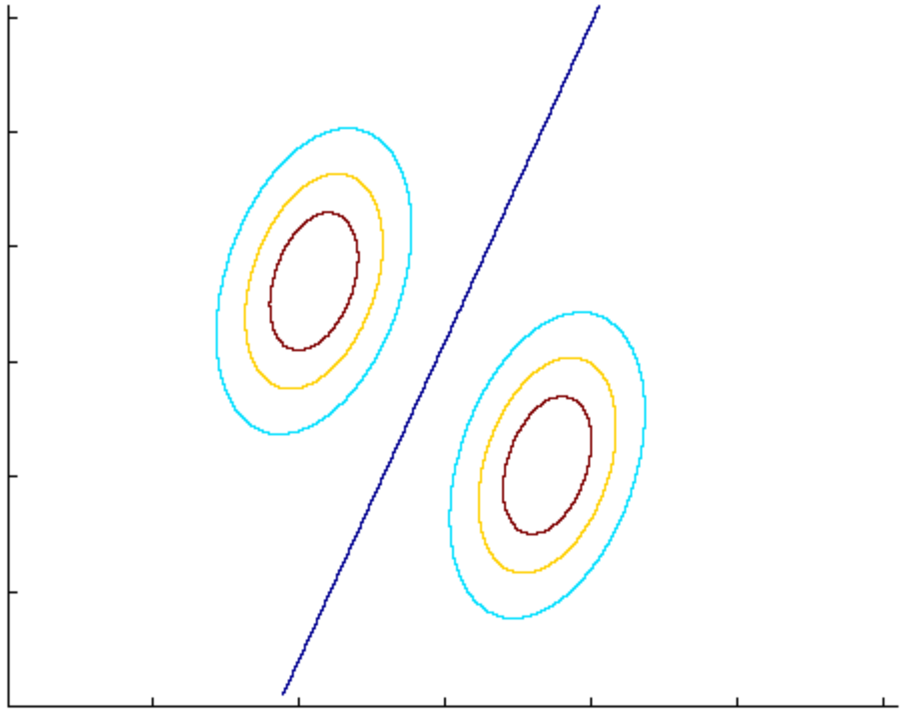
$$\text{Correlation}: \mathrm{Corr}(X_i, X_j) \equiv \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

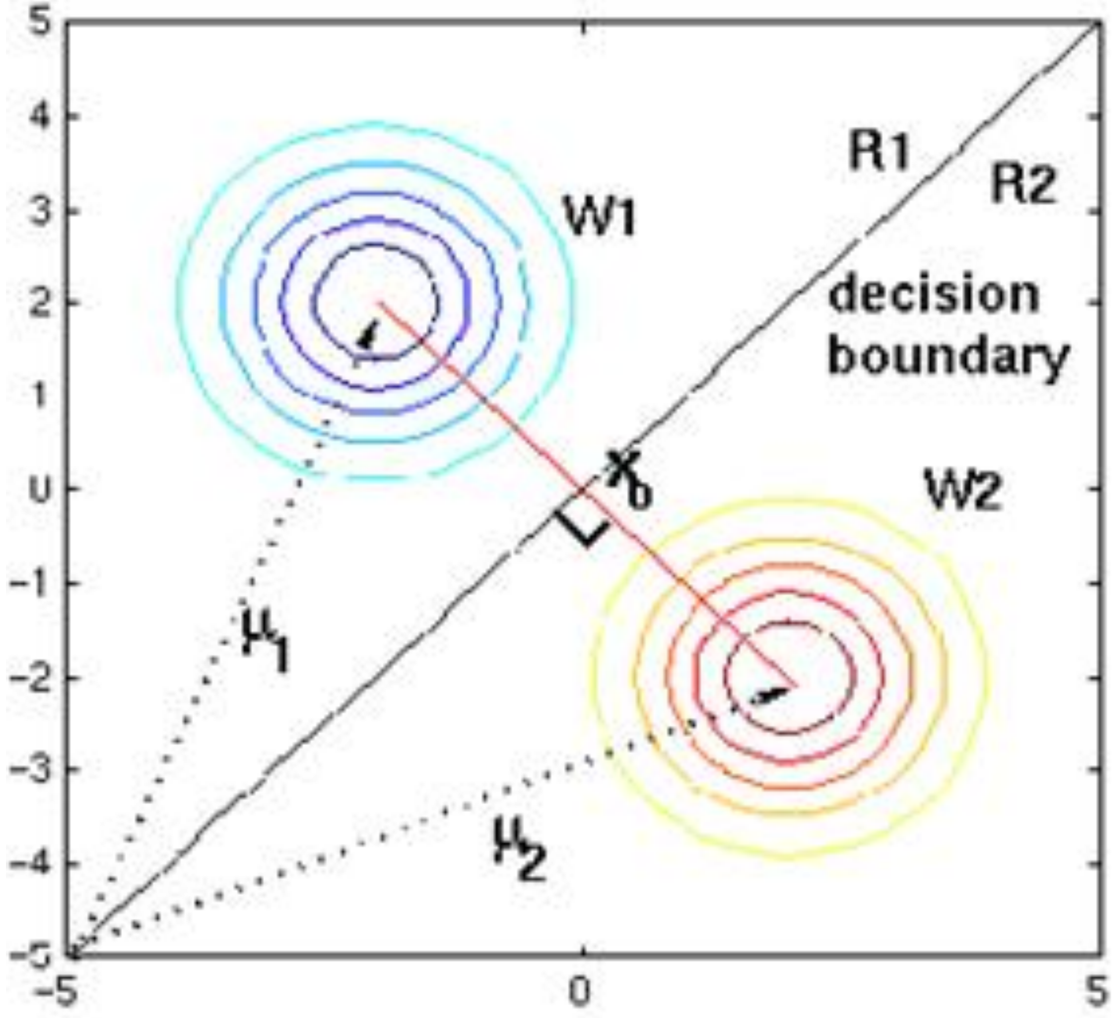
Bivariate Normal

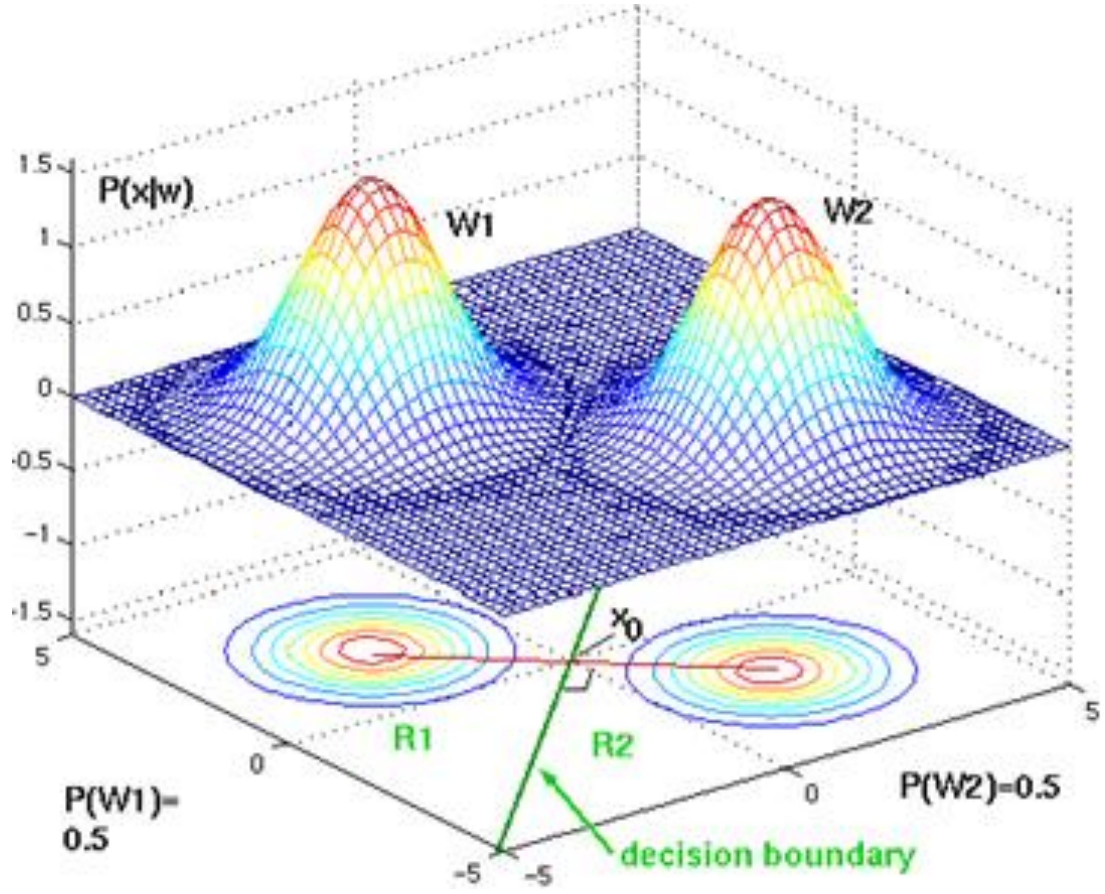
How do changes in our underlying expectations of the generating models impact our decisions to classify?

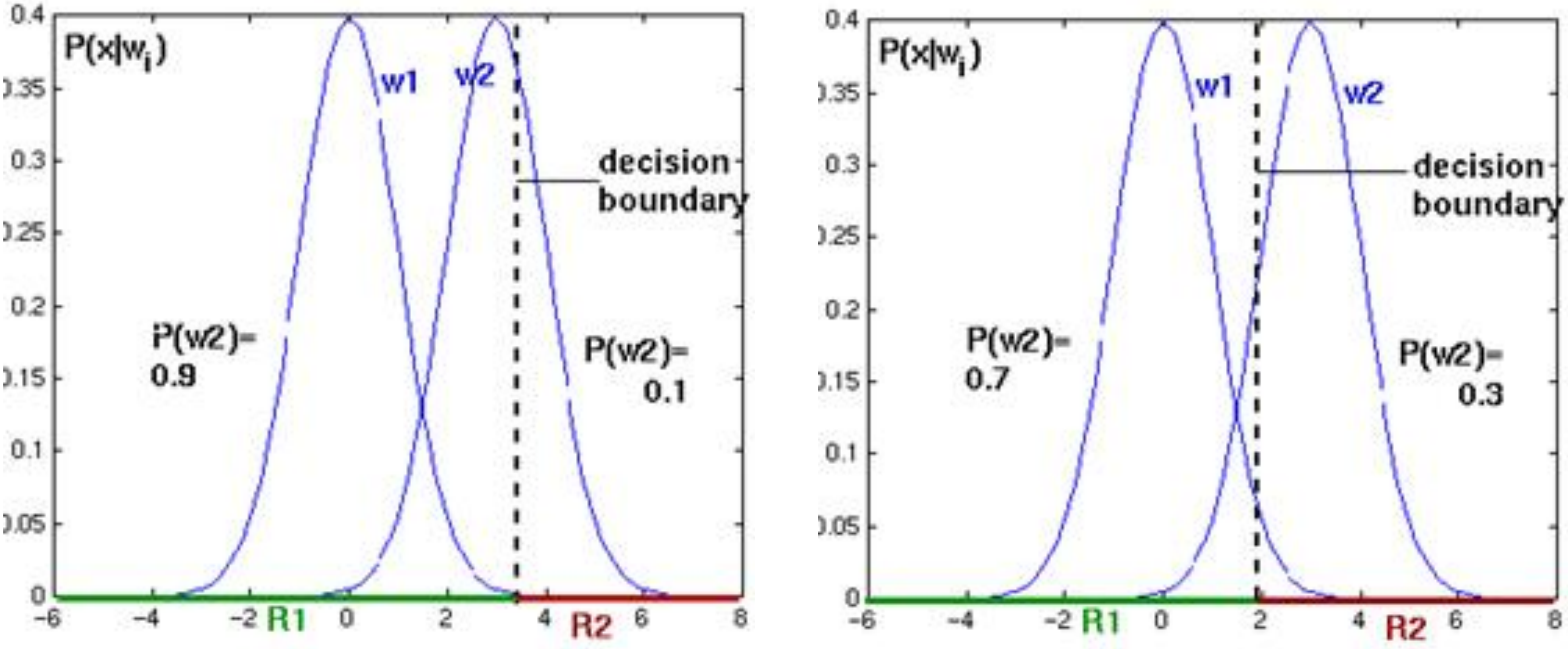


Common Covariance




R1
R2
W1
decision
boundary
W2
x_0
μ_1
μ_2


P(x|w)
W1
W2
x_0
R1
R2
P(W1)=
0.5
P(W2)=0.5
decision boundary



P(x|w_i)
w1
w2
decision boundary
P(w2)= 0.9
P(w2)= 0.1
R1
R2
P(x|w_i)
w1
w2
decision boundary
P(w2)= 0.7
P(w2)= 0.3
R1
R2

Bayesian estimators

It seems much more natural to attempt to make statements about which parameter values are likely given the data you have collected

To put this on a rigorous probabilistic footing we want to make statements about the probability (density) of any particular parameter value given our data

Bayes theorem:

Posterior

Prior

Likelihood

$$P(\theta \mid D) = \frac{P(\theta)P(D \mid \theta)}{P(D)}$$

Normalising constant

Bayes estimators

The single most important conceptual difference between Bayesian statistics and frequentist statistics is the notion that the parameters you are interested in are themselves random variables

This notion is encapsulated in the use of a subjective prior for your parameters

Remember that to construct a confidence interval we have to define the set of possible parameter values

A prior does the same thing, but also gives a weight to different values

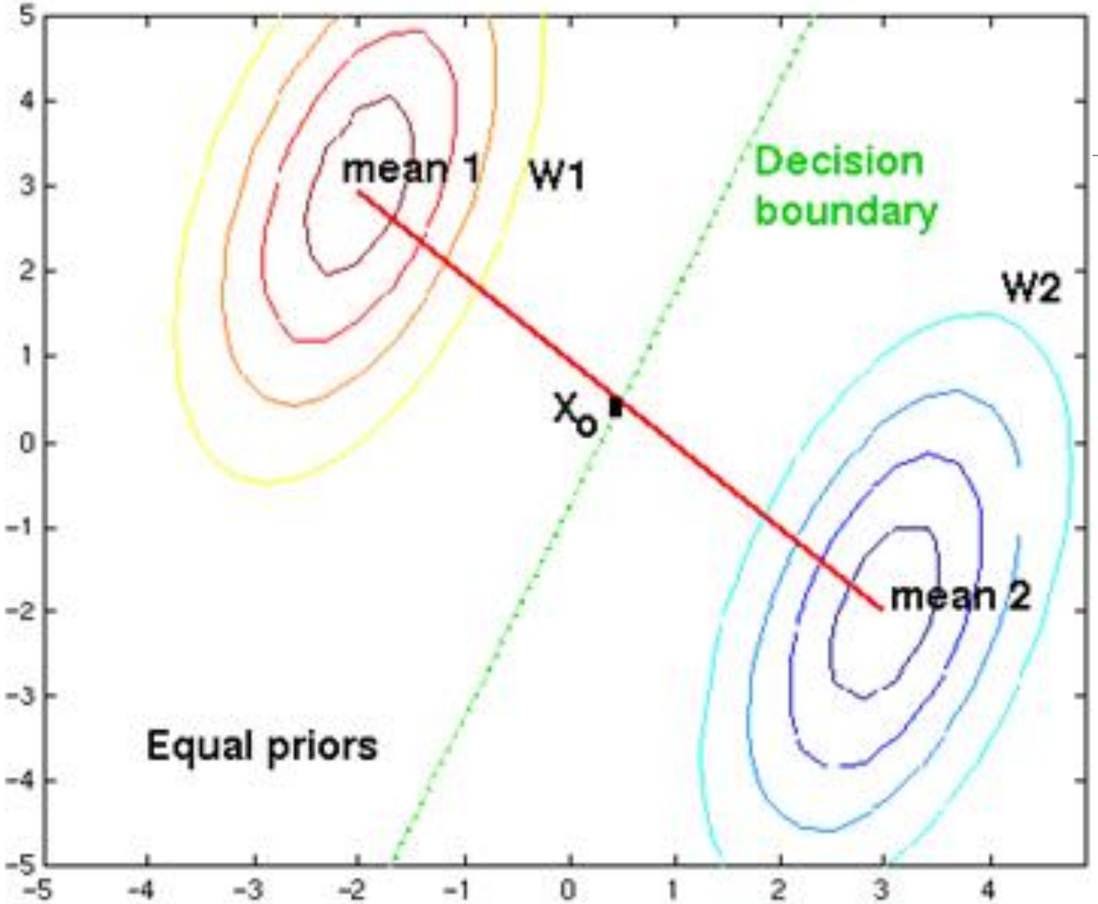
Example: Coin Tossing

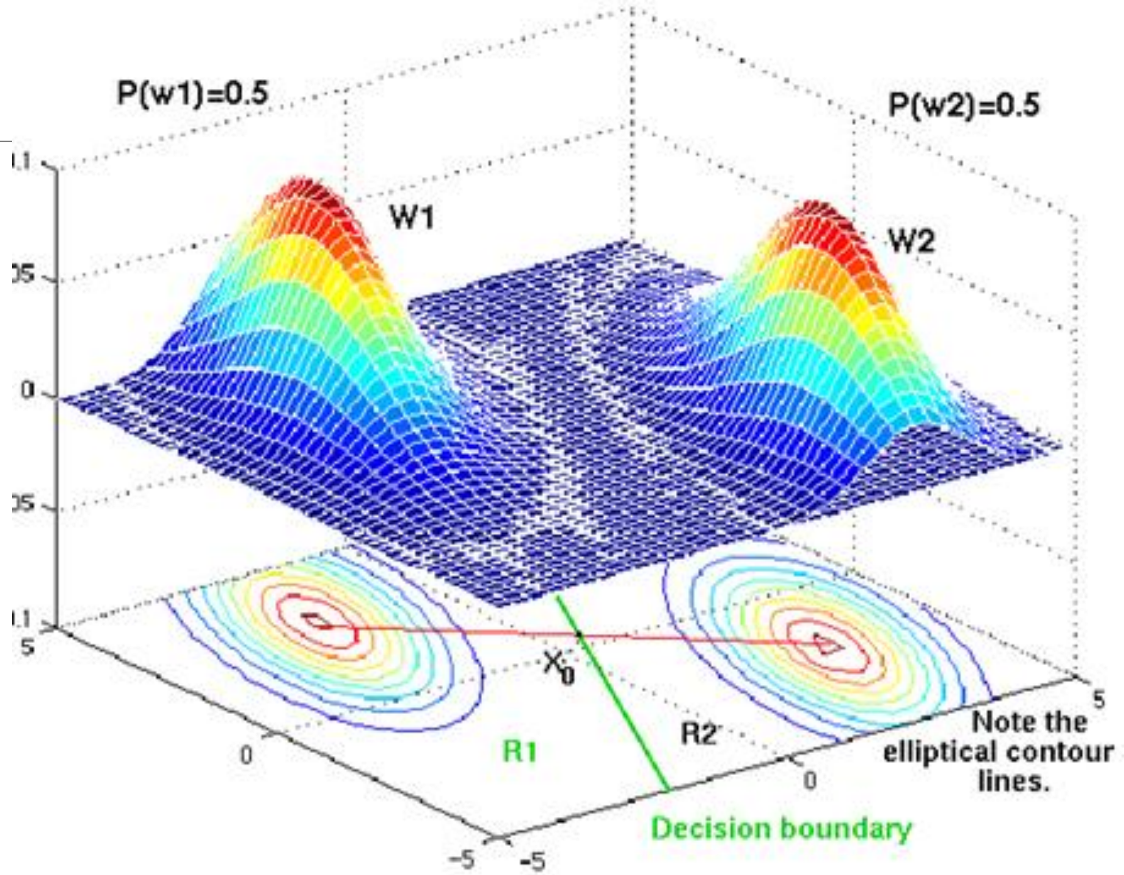
I toss a coin twice and observe two heads

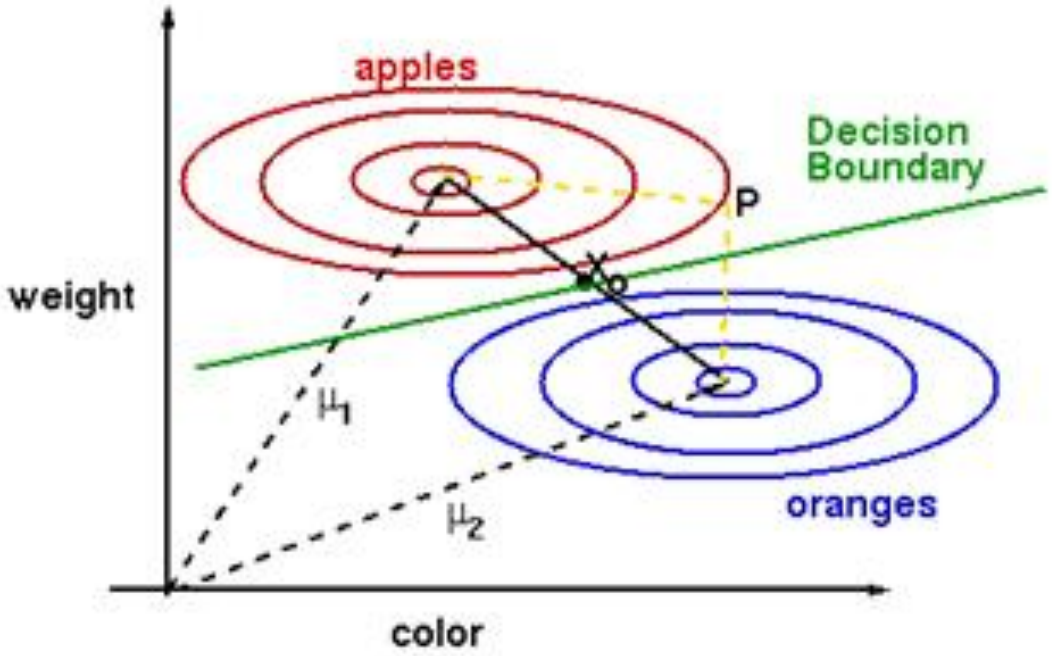
I want to perform inference about the probability of obtaining a head on a single throw for the coin in question

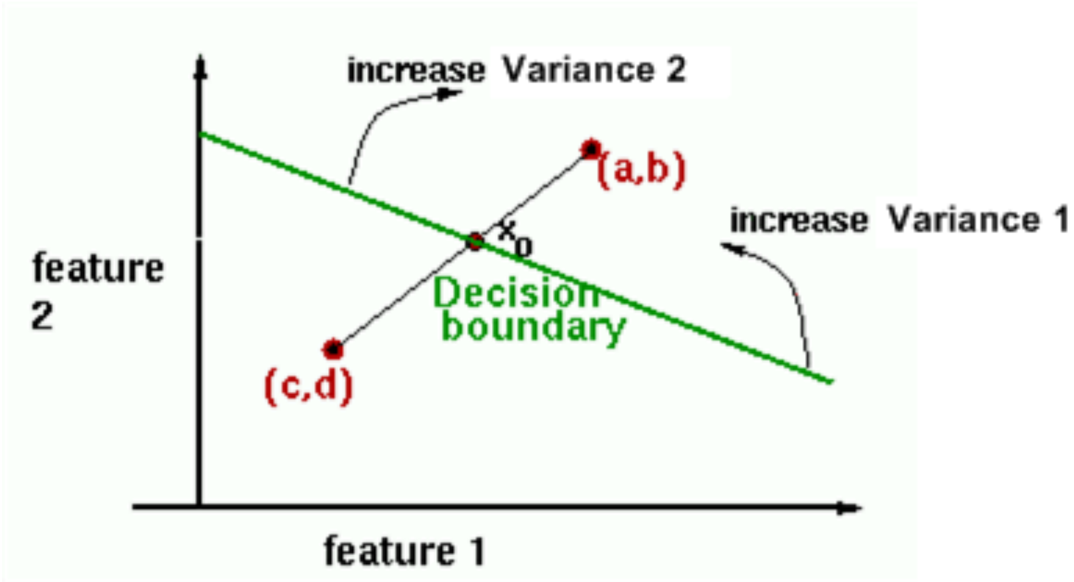
The point estimate/MLE for the probability is 1.0 – yet I have a very strong prior belief that the answer is 0.5

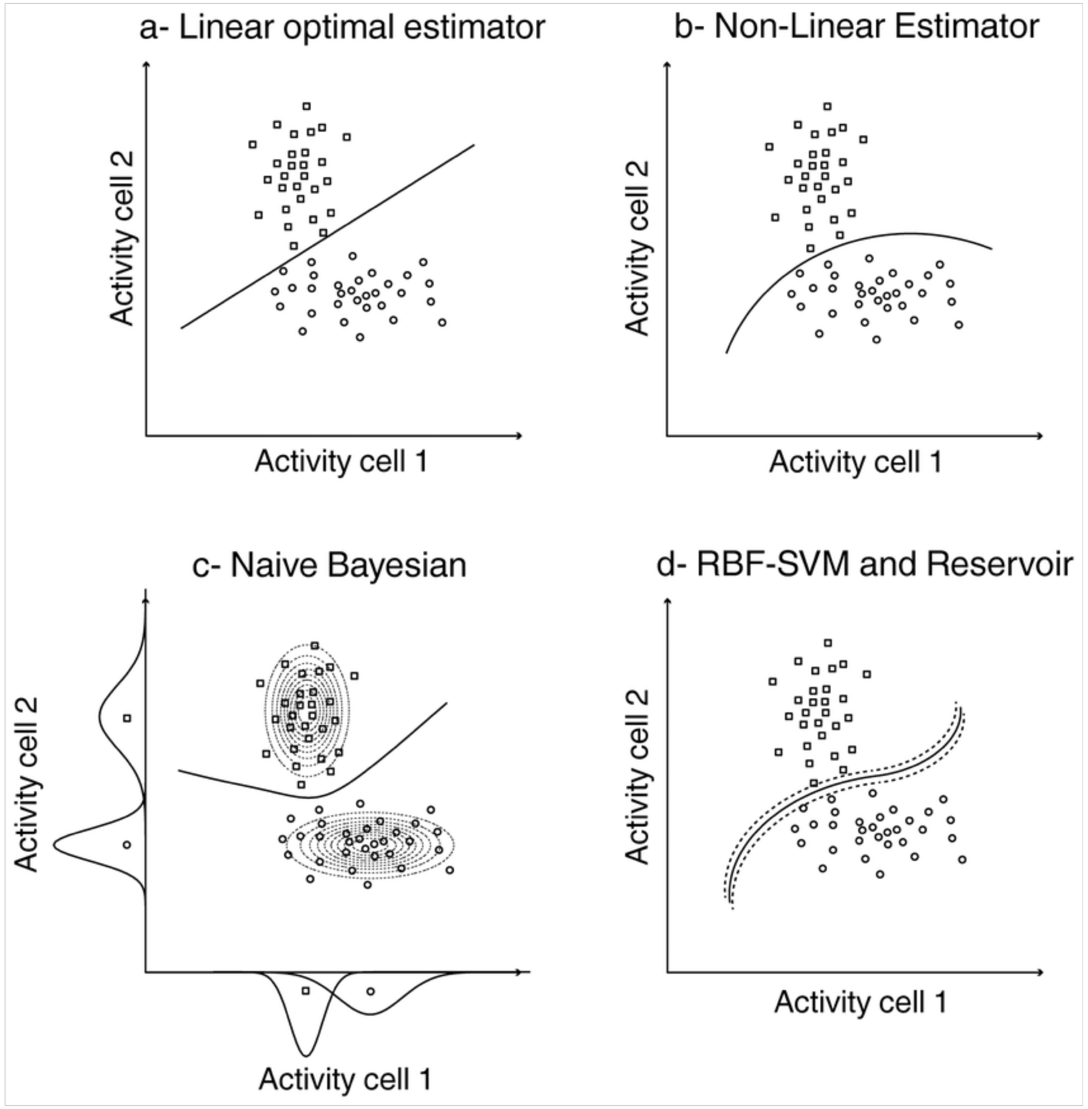
Bayesian statistics forces the researcher to be explicit about prior beliefs but, in return, can be very specific about what information has been gained by performing the experiment


mean 1
W1
Decision
boundary
W2
X0
mean 2
Equal priors


P(w1)=0.5
P(w2)=0.5
W1
W2
X0
R1
R2
Note the
elliptical contour
lines.
Decision boundary


apples
Decision
Boundary
P
x_o
weight
μ_1
μ_2
oranges
color


increase Variance 2
(a,b)
increase Variance 1
x_o
feature
2
Decision
boundary
(c,d)
feature 1



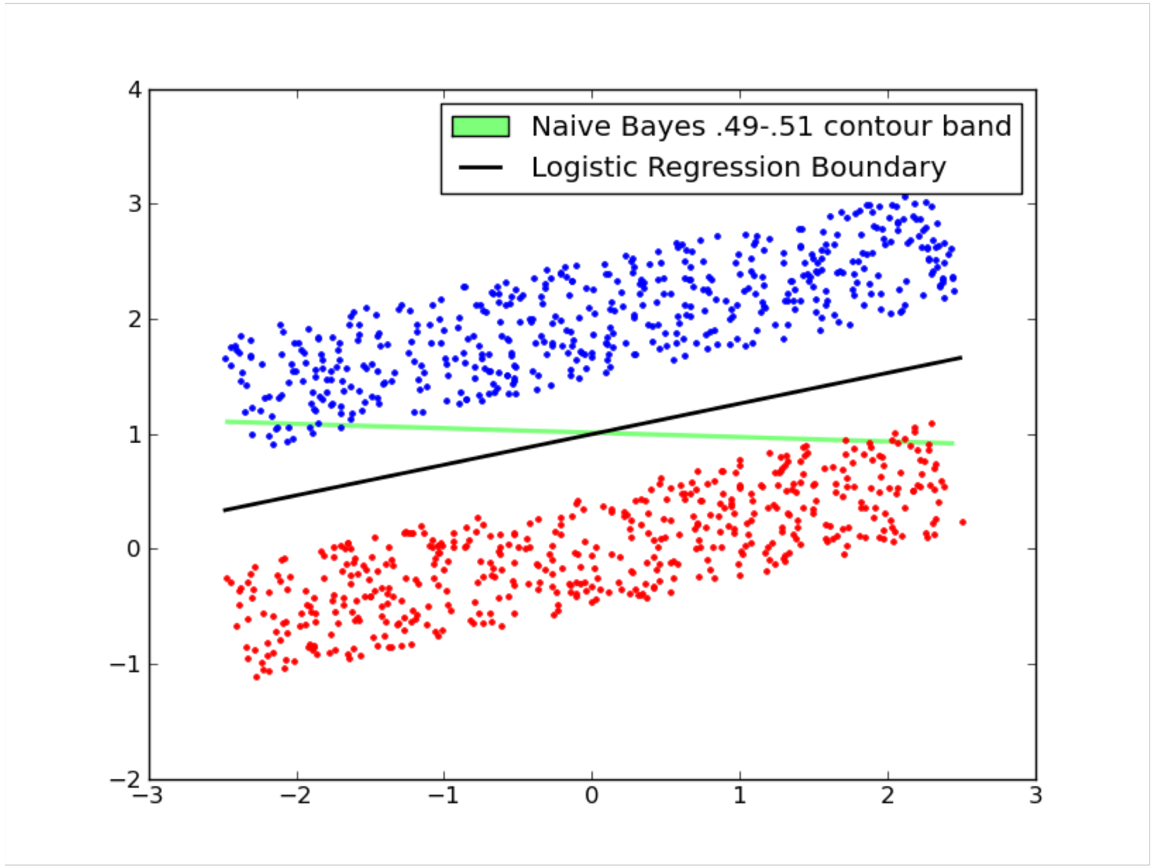
MLP vs MAP

We've done maximum likelihood before (Linear Regression)

- Constraints?
  - Gaussian distributed data

We've done maximum a posteriori before (Naïve Bayes)

- Constraints?
  - All features are independent (necessary computationally)
  - No cross correlation
- What if we calculate it exactly?
  - Bayesian estimator



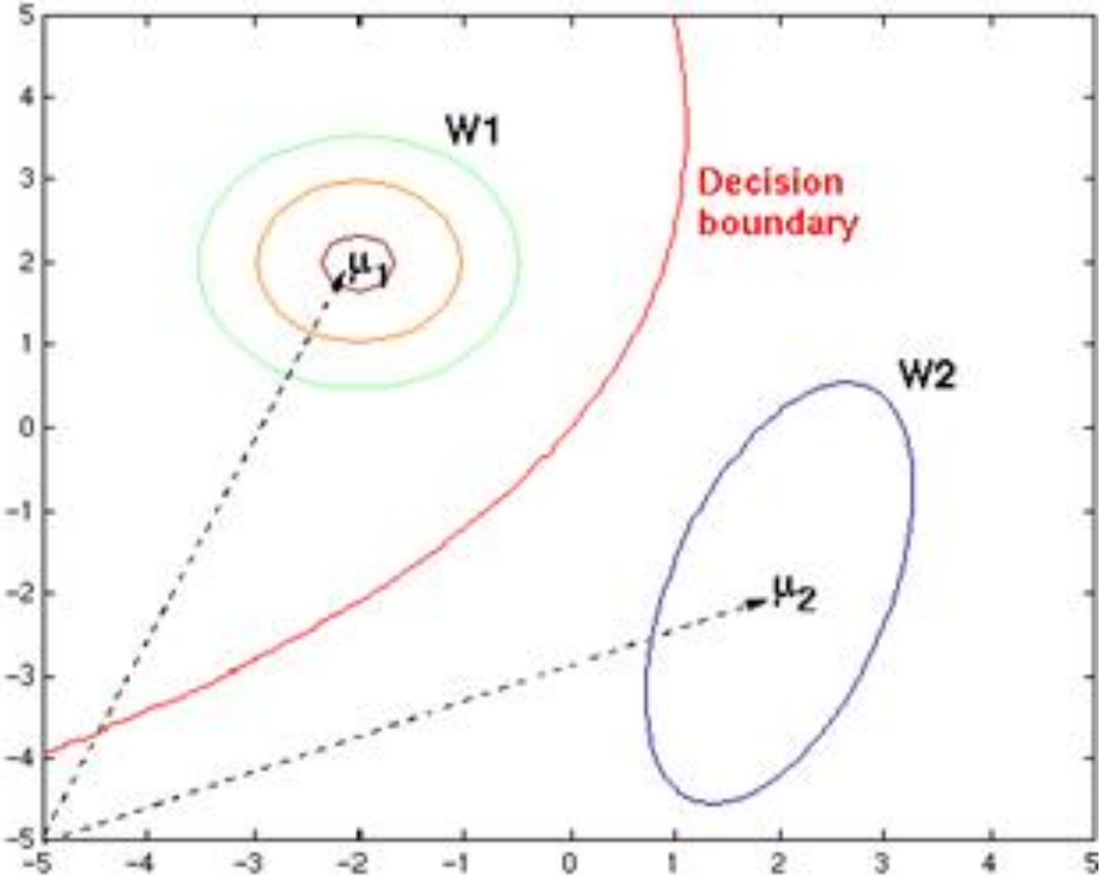
MLP vs MAP

We've done maximum likelihood before (Linear Regression)

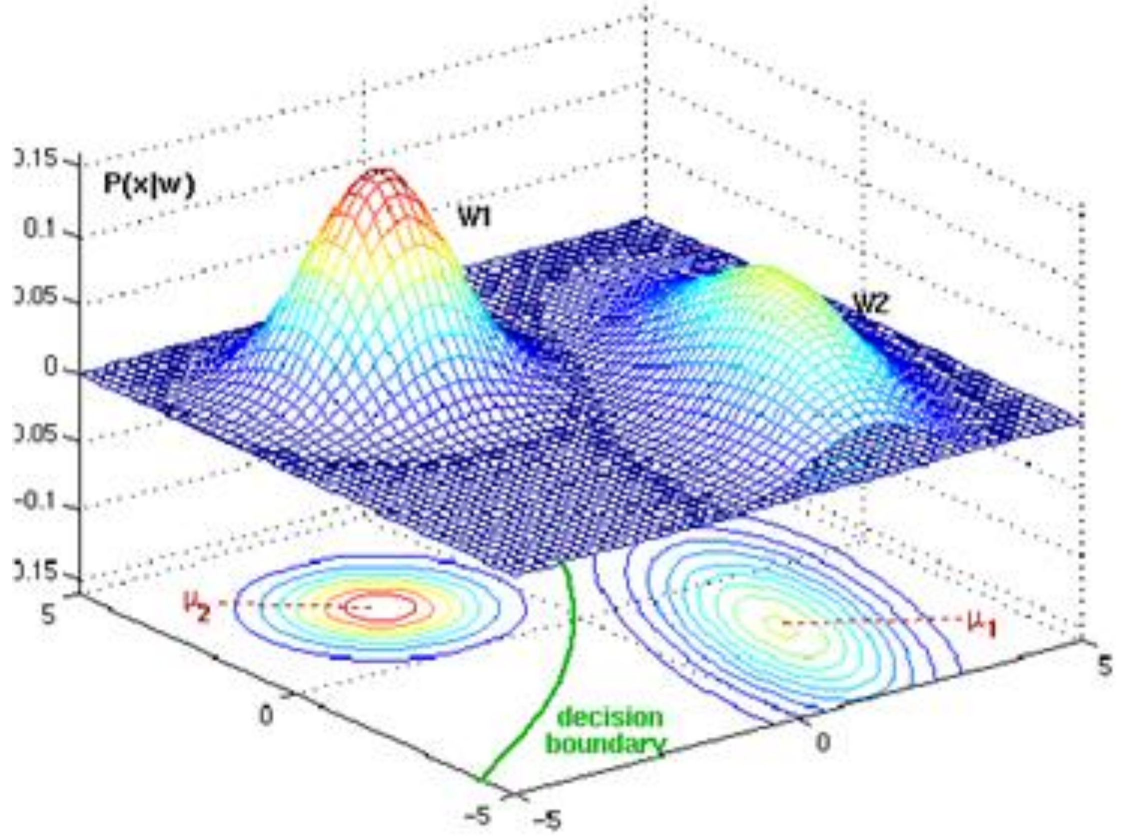
- Constraints?
  - Gaussian distributed data

We've done maximum a posteriori before (Naïve Bayes)

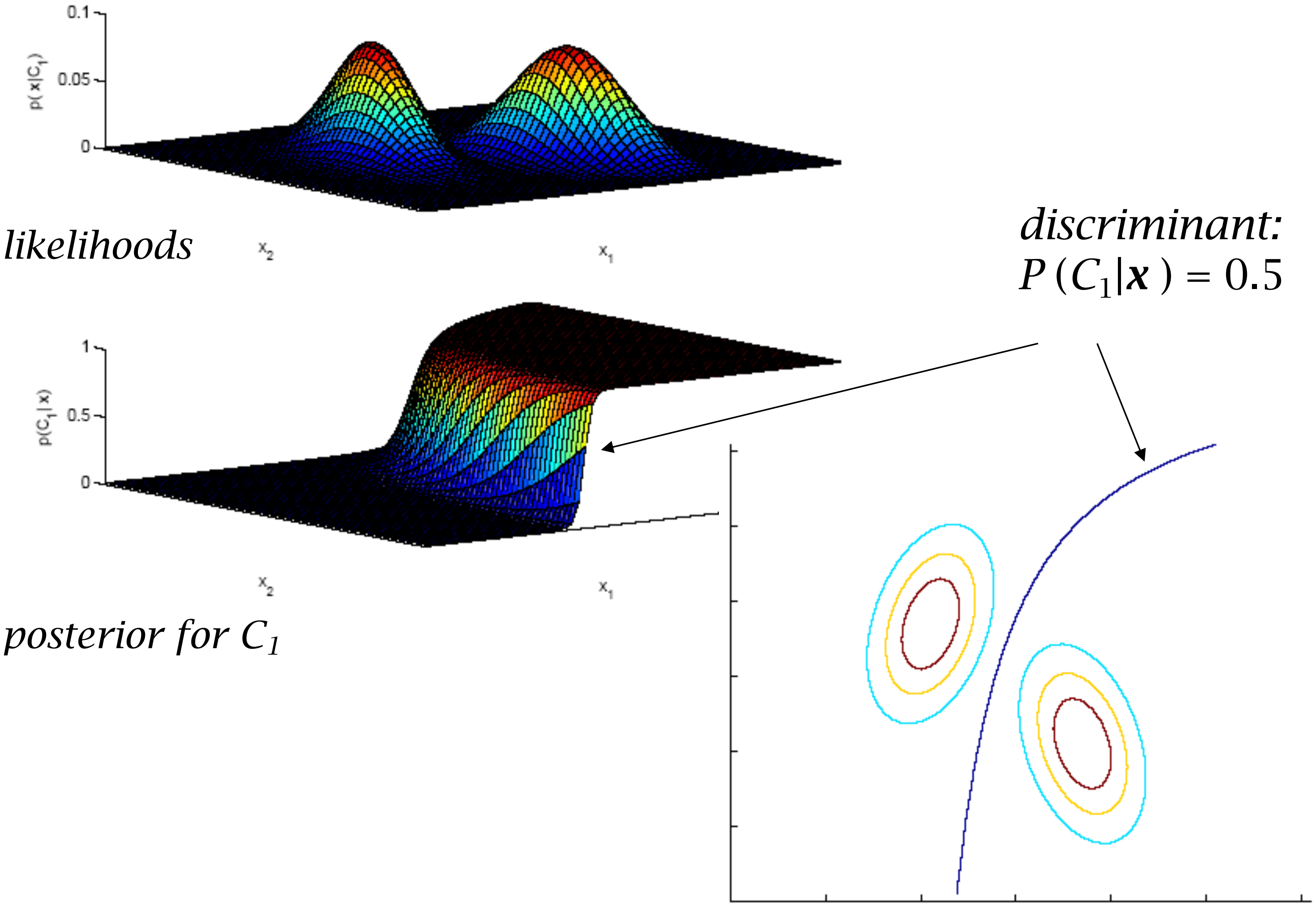
- Constraints?
  - All features are independent (necessary computationally)
  - No cross correlation
- What if we calculate it exactly?
  - Bayesian estimator



W1
Decision boundary
W2
μ_1
μ_2



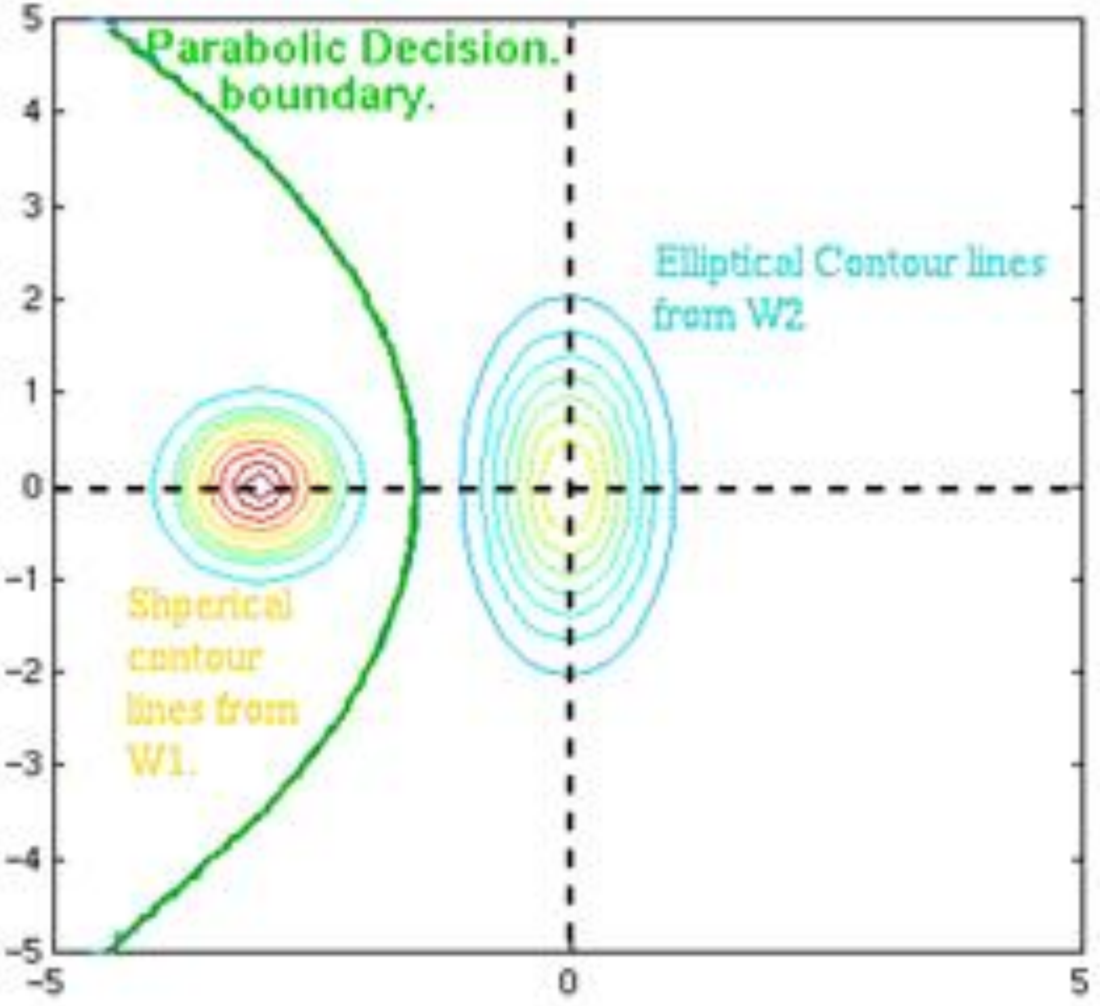
P(x|w)
W1
W2
μ_2
μ_1
decision boundary



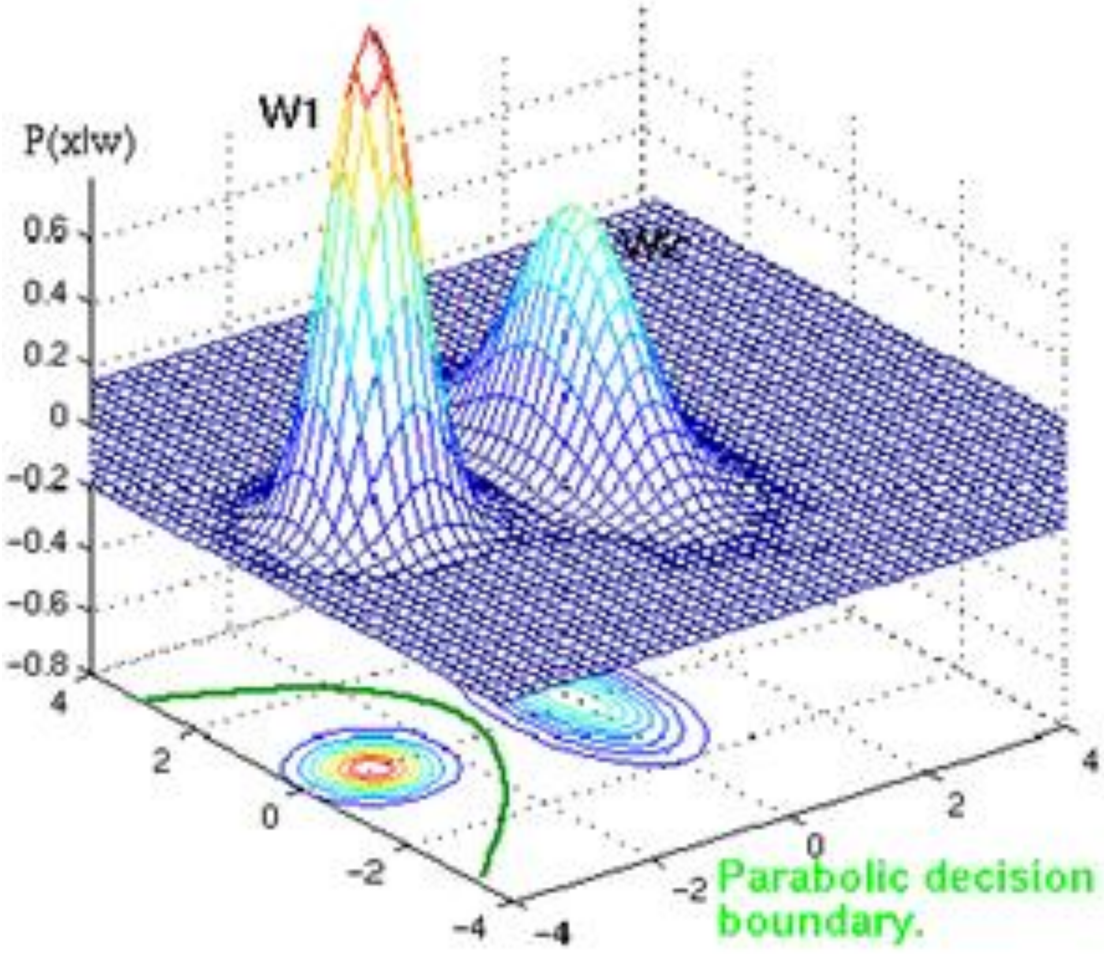
likelihoods

posterior for C_1

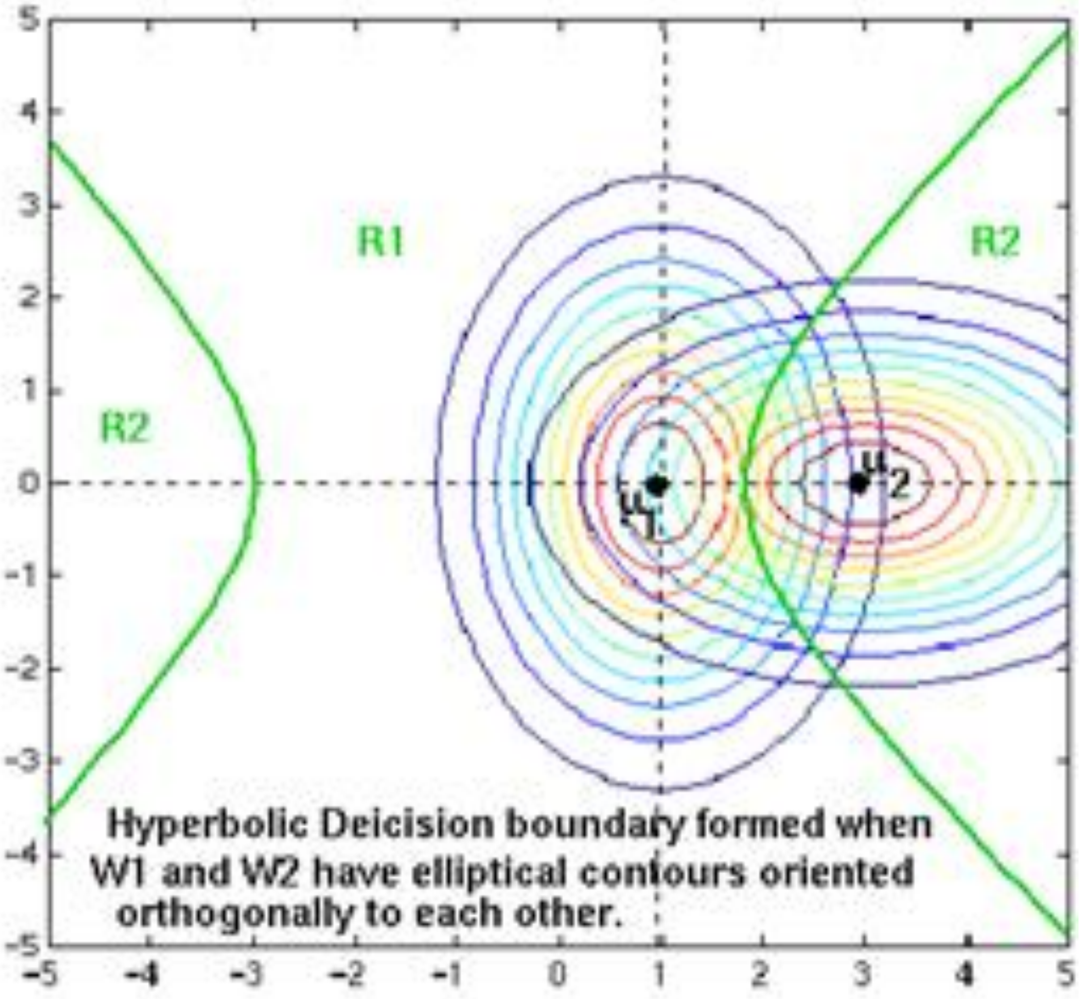
discriminant:
$P(C_1|\boldsymbol{x}) = 0.5$

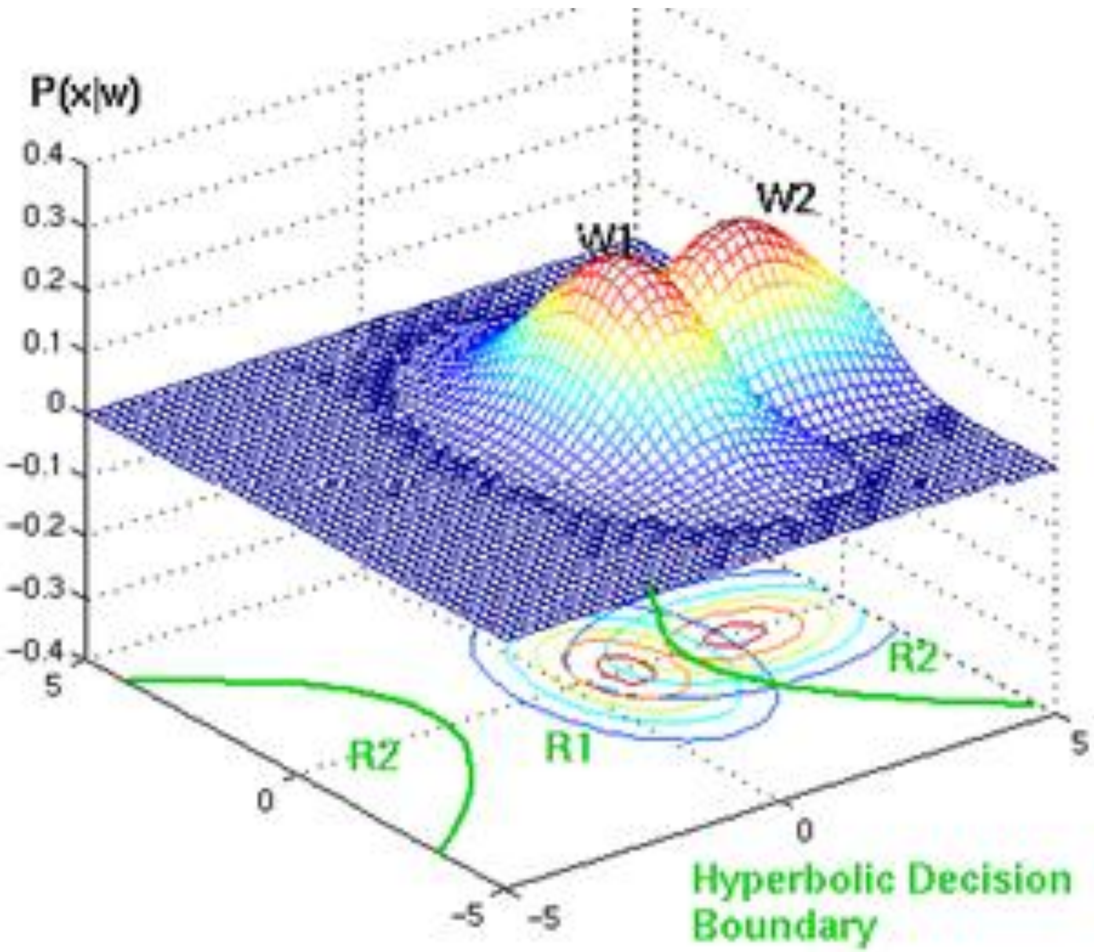


Parabolic Decision. boundary.
Elliptical Contour lines from W2
Shperical contour lines from W1.



W1
P(x|w)
Parabolic decision boundary.


R1
R2
R2
μ_1
μ_2
Hyperbolic Deicision boundary formed when
W1 and W2 have elliptical contours oriented
orthogonally to each other.


P(x|w)
W1
W2
R2
R1
R2
Hyperbolic Decision
Boundary