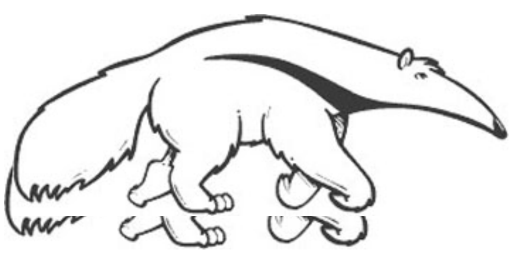
Machine Learning and Data Mining

Bayes Classifiers; Naïve Bayes

Prof. Alexander Ihler

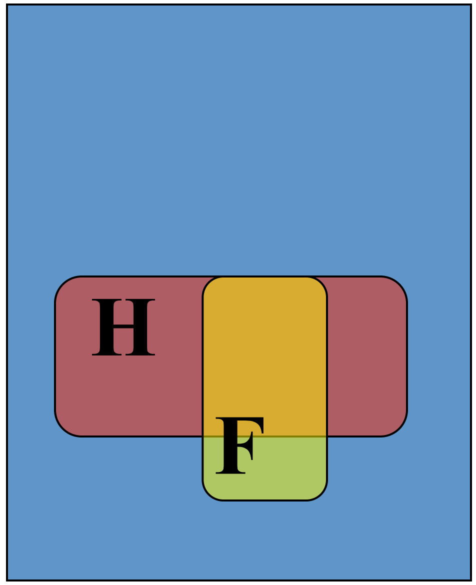
Fall 2012



BREN:ICS
INFORMATION AND COMPUTER SCIENCES
UNIVERSITY of CALIFORNIA IRVINE

Bayes rule

- Two events: headache, flu
- p(H) = 1/10
- p(F) = 1/40
- p(H|F) = 1/2



- You wake up with a headache – what is the chance that you have the flu?

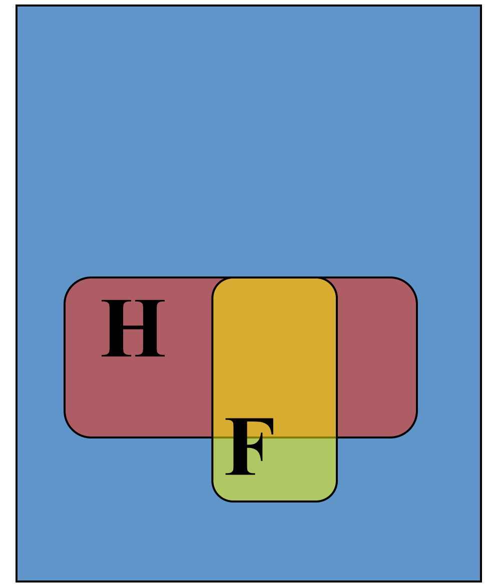
Example from Andrew Moore's slides

Bayes rule

- Two events: headache, flu
- p(H) = 1/10
- p(F) = 1/40
- p(H|F) = 1/2

- P(H & F) = ?

- P(F|H) = ?



Example from Andrew Moore's slides

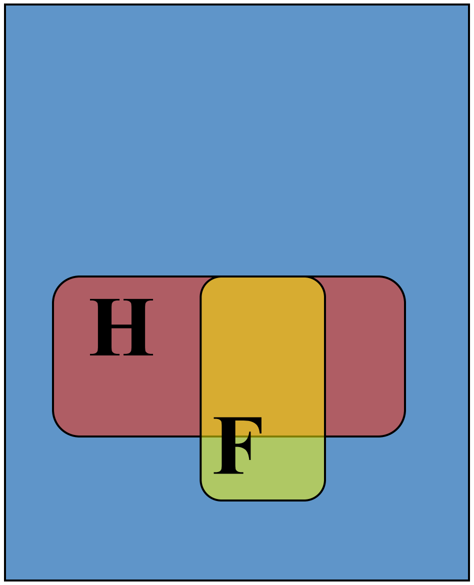
Bayes rule

- Two events: headache, flu
- $p(H) = 1/10$
- $p(F) = 1/40$
- $p(H|F) = 1/2$

- $P(H \& F) = p(F)\, p(H|F)$

$$= (1/2) * (1/40) = 1/80$$

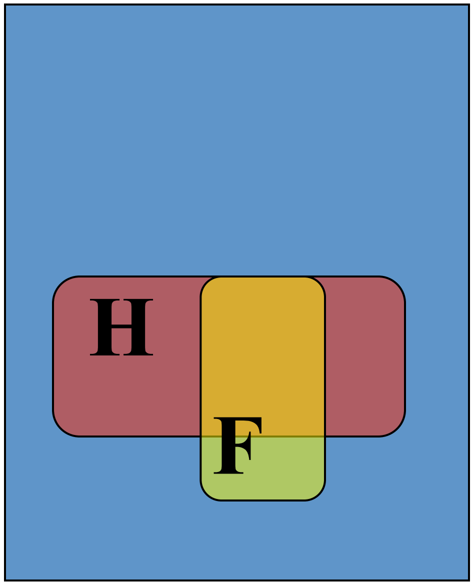
- $P(F|H) = ?$



Example from Andrew Moore's slides

Bayes rule

- Two events: headache, flu
- p(H) = 1/10
- p(F) = 1/40
- p(H|F) = 1/2



- P(H & F) = p(F) p(H|F)
  = (1/2) * (1/40) = 1/80
- P(F|H) = p(H & F) / p(H)
  = (1/80) / (1/10) = 1/8

Example from Andrew Moore's slides

Classification and probability

- Suppose we want to model the data
- Prior probability of each class, p(c)
  - E.g., fraction of emails that are spam
- Distribution of features given the class, p(x | c)
  - How likely are we to see “x” in spam?
- Joint distribution

$$p(c|x)p(x) = p(x,c) = p(x|c)p(c)$$

- Bayes Rule:

$$\Rightarrow \quad p(c|x) = p(x|c)p(c)/p(x)$$
$$= \frac{p(x|c)p(c)}{\sum_c p(x|c)p(c)}$$

Bayes classifiers

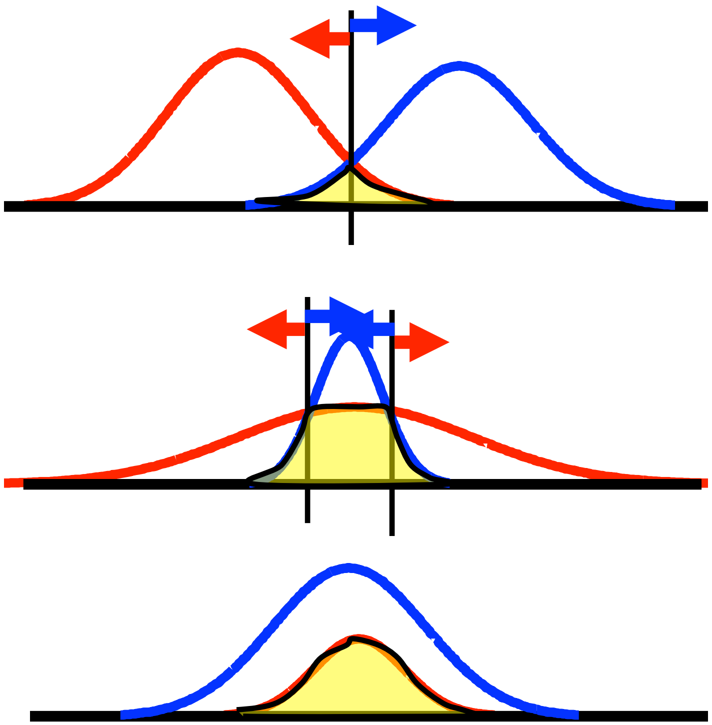
- Estimate p(c)=[p(C=0), p(C=1) …]
- Estimate p(x|C=c) for each class C
- Calculate p(C=c|x) using Bayes rule
- Choose the most likely class c

Bayes rule:

$p(C=c \mid x) = p(x|C=c) * p(C=c) / p(x)$

Rule of total probability:

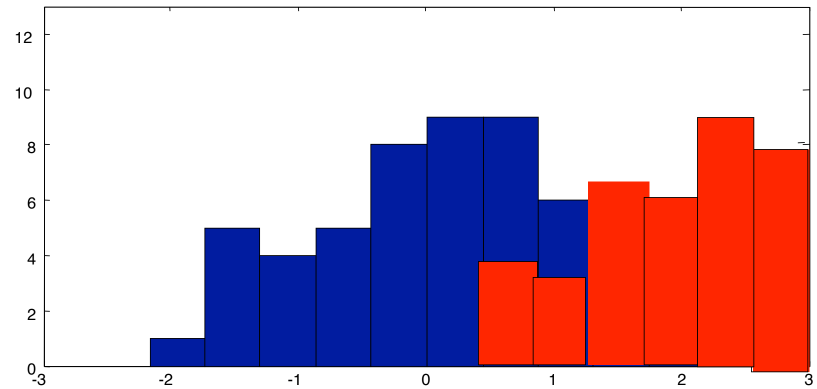
$p(x) = p(x|C=0)p(C=0) + p(x|C=1)p(C=1) + \ldots$



Bayes classifiers

- Learn "class conditional" models
  - Estimate a probability model for each class
- Training data
  - Split by class
  - $D_c = \{ x^{(j)} : y^{(j)} = c \}$
- Estimate $p(x \mid y=c)$ using D_c

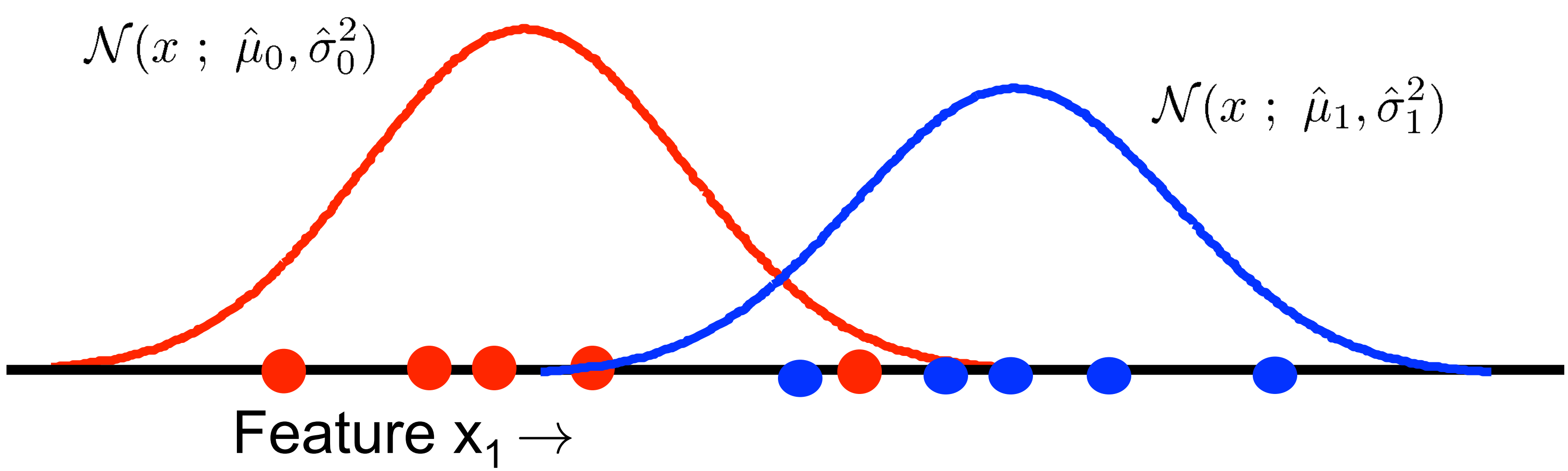
- Can use any density estimate we'd like
  - Histogram
  - Gaussian
  - …



Gaussian models

- Estimate parameters of the Gaussians from the data

$$\alpha = \frac{m_1}{m} = \hat{p}(y = c_1) \qquad \hat{\mu} = \frac{1}{m} \sum_j x^{(j)} \qquad \hat{\sigma}^2 = \frac{1}{m} \sum_j (x^{(j)} - \mu)^2$$



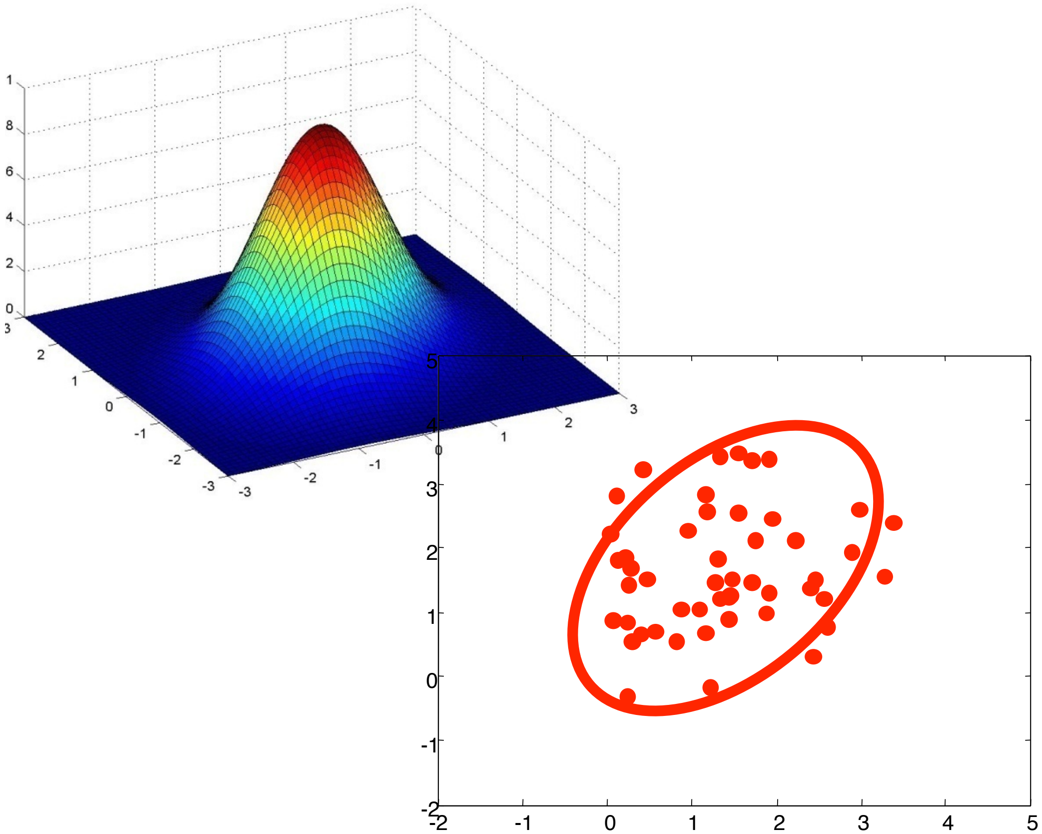
Multivariate Gaussian models

- Similar to univariate case

$$\mathcal{N}(\underline{x}\ ;\ \underline{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2}} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}) \right\}$$

μ = length-d column vector
Σ = d x d matrix

$|\Sigma|$ = matrix determinant



Maximum likelihood estimate:

$$\hat{\mu} = \frac{1}{m} \sum_j \underline{x}^{(j)}$$

$$\hat{\Sigma} = \frac{1}{m} \sum_j (\underline{x}^{(j)} - \underline{\hat{\mu}})^T (\underline{x}^{(j)} - \underline{\hat{\mu}})$$

Joint distributions

- Make a truth table of all combinations of values

A,	B,	C
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

Joint distributions

- Make a truth table of all combinations of values
- For each combination of values, determine how probable it is
- Total probability must sum to one
- How many values did we specify?

A,	B,	C	p(.)
0	0	0	0.50
0	0	1	0.05
0	1	0	0.01
0	1	1	0.10
1	0	0	0.04
1	0	1	0.15
1	1	0	0.05
1	1	1	0.10

Overfitting and density estimation

- Estimate probabilities from the data
  - E.g., how many times (what fraction) did each outcome occur?

- M data << 2^N parameters?

- What about the zeros?
  - We learn that certain combinations are impossible?
  - What if we see these later in test data?

- Overfitting!

A	B	C	p(.)
0	0	0	4/10
0	0	1	1/10
0	1	0	0/10
0	1	1	0/10
1	0	0	1/10
1	0	1	2/10
1	1	0	1/10
1	1	1	1/10

Overfitting and density estimation

A	B	C	p(.)
0	0	0	4/10
0	0	1	1/10
0	1	0	0/10
0	1	1	0/10
1	0	0	1/10
1	0	1	2/10
1	1	0	1/10
1	1	1	1/10

- Estimate probabilities from the data
  - E.g., how many times (what fraction) did each outcome occur?
- M data << 2^N parameters?
- What about the zeros?
  - We learn that certain combinations are impossible?
  - What if we see these later in test data?
- One option: regularize $\hat{p}(a,b,c) \propto (N_{abc} + \alpha)$
- Normalize to make sure values sum to one…

Overfitting and density estimation

- Another option: reduce the model complexity
  - E.g., assume that features are independent of one another

- Independence:
- $p(x,y) = p(x)\, p(y)$

- $p(x1, x2, \ldots xN) = p(x1)\, p(x2) \ldots p(xN)$
- Only need to estimate each individually

A,		B,		C	
0	**.4**	**0**	**.7**	**0**	**.1**
1	**.6**	**1**	**.3**	**1**	**.9**

A,	B,	C	p(.)
0	**0**	**0**	**4/10**
0	**0**	**1**	**1/10**
0	**1**	**0**	**0/10**
0	**1**	**1**	**0/10**
1	**0**	**0**	**1/10**
1	**0**	**1**	**2/10**
1	**1**	**0**	**1/10**
1	**1**	**1**	**1/10**

Conditional independence

- Ex: cavity, toothache, "catch"
  - Toothache and "catch" are not independent
  - But probably independent given cavity=1 or cavity=0

- Conditional independence:
  - $p(y,z \mid x) = p(y|x)\, p(z|x)$ $\qquad y \perp z \mid x$
  - z only depends (directly) on x, not y
  - z and y are coupled through x

Conditional independence

- Fully general distribution:
  - $p(x,y,z) = p(x)\, p(y|x)\, p(z|x,y)$
  - (mx*my*mz - 1) free parameters

- Conditionally independent, $y \perp z \mid x$
  - $p(x,y,z) = p(x)\, p(y|x)\, p(z|x)$
  - (mx-1) + (my-1)*mx + (mz-1)*mx
  - Much fewer

- Ex: mx=my=mz=10
  - Arbitrary joint dist = 999 free parameters
  - Conditionally independent dist = 189 parameters

Naïve Bayes models

- Suppose we have some variable y to predict
  - Ex: risk of auto accident
- We have *many* co-observed vars $\mathbf{x}=[x_1 \ldots x_m]$
  - Age, income, education, zip code, …
- Want to learn $p(y \mid x_1 \ldots x_m)$, to predict y
- Arbitrary distribution: $O(d^{m+1})$ values!

- Naïve Bayes:
  - $p(y|\mathbf{x}) = p(\mathbf{x}|y)\, p(y) / p(\mathbf{x})$; $p(\mathbf{x}|y) = \prod_i p(x_i|y)$
  - Covariates are independent given "cause"

- May not be a good model of the data
  - Doesn't capture correlations in x's
  - Can't capture some dependencies
- But in practice it often does quite well!

Naïve Bayes Models for Spam

- y $\in$ {spam, not spam}
- X = observed words in email
  - Ex: [“the” … “probabilistic” … “lottery”…]
  - “1” if word appears; “0” if not
- 1000’s of possible words: 2^{1000s} parameters?
- # of atoms in the universe: $\sim 2^{270}$…

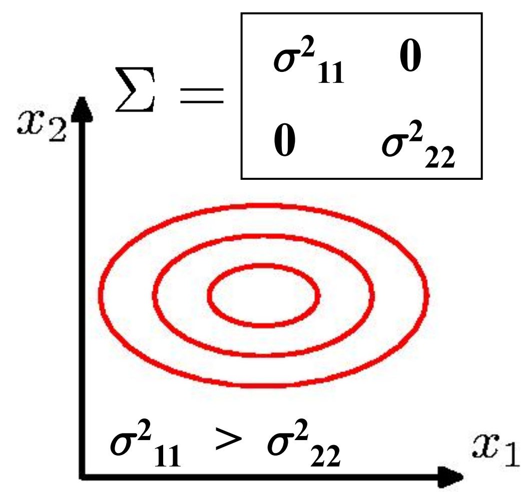
- Model words *given* email type as independent
- Some words more likely for spam (“lottery”)
- Some more likely for real (“probabilistic”)
- Only 1000’s of parameters now…

Naïve Bayes Gaussian models

$$p(x_1) = \frac{1}{Z} \exp\left\{ -\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 \right\} \qquad p(x_2) = \frac{1}{Z_2} \exp\left\{ -\frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \right\}$$

$$p(x_1)p(x_2) = \frac{1}{Z_1 Z_2} \exp\left\{ -\frac{1}{2}(\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}) \right\}$$

$$\underline{\mu} = [\mu_1 \;\; \mu_2]$$
$$\Sigma = \mathrm{diag}(\sigma_1^2 \;,\; \sigma_2^2)$$



(b)

You should know…

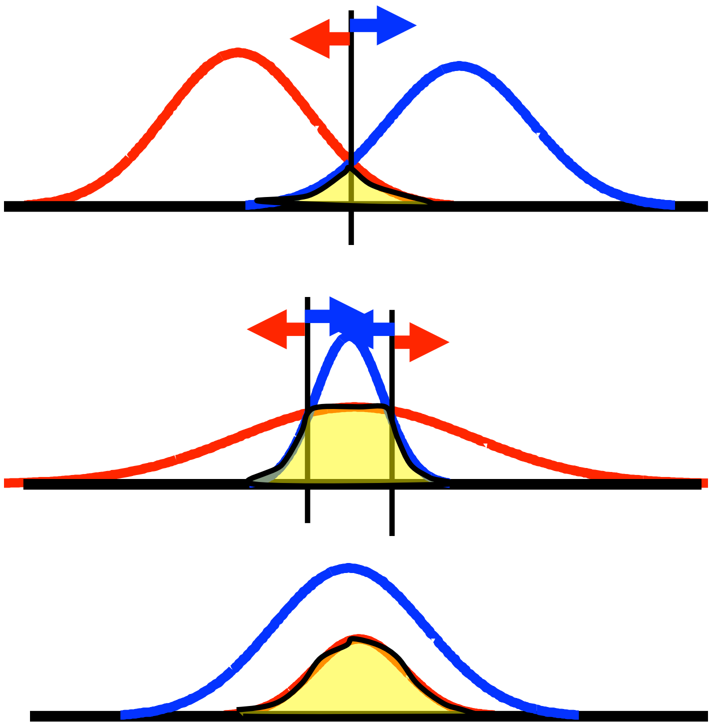
- Bayes rule; $p(c|x)$
- Bayes classifiers
  - Learn $p(x|c=C)$, $p(c=C)$
- Naïve Bayes classifiers
  - Assume $p(x|c=C) = p(x_1|c=C)\, p(x_2|c=C) \ldots$

- Maximum likelihood estimators
  - Discrete variables
  - Gaussian variables
  - Overfitting; simplifying assumptions or regularization

Gaussian models

- "Bayes optimal" decision
  - Choose most likely class
- Decision boundary
  - Places where probabilities equal

- What shape is the boundary?



Gaussian models

- Bayes optimal decision boundary
  - p(y=0 | x) = p(y=1 | x)
  - Transition point between p(y=0|x) >/< p(y=1|x)
- Assume Gaussian models with equal covariances

$$\mathcal{N}(\underline{x}\ ;\ \underline{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2}} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}) \right\}$$

$$0 \underset{>}{<} \log \frac{p(x|y=0)}{p(x|y=1)} \frac{p(y=0)}{p(y=1)} = \log \frac{p(y=0)}{p(y=1)} +$$

$$-.5(x\Sigma^{-1}x - 2\mu_0^T \Sigma^{-1} x + \mu_0^T \Sigma^{-1} \mu_0)$$

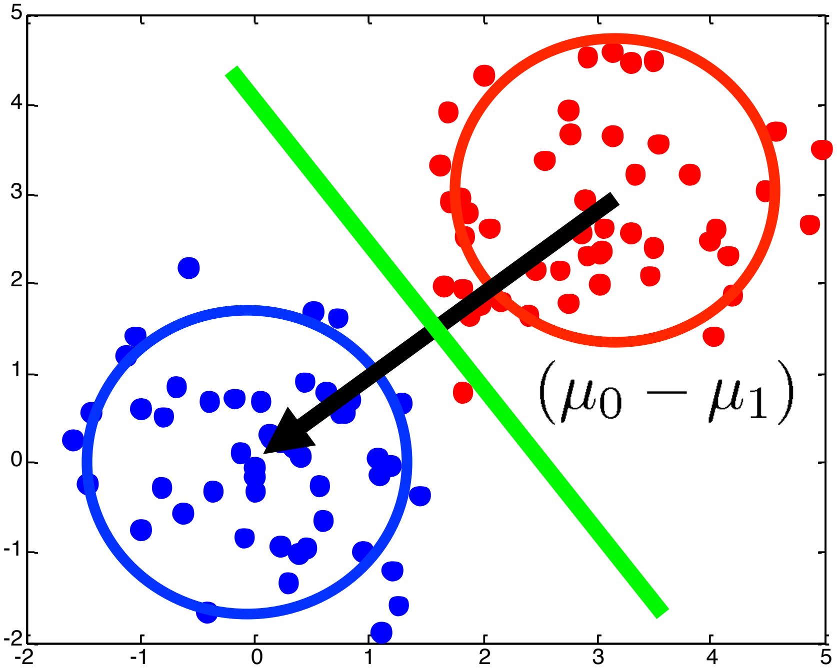
$$+.5(x\Sigma^{-1}x - 2\mu_1^T \Sigma^{-1} x + \mu_1^T \Sigma^{-1} \mu_1)$$

$$= (\mu_0 - \mu_1)^T \Sigma^{-1} x + constants$$

Gaussian example

- Spherical covariance: $\Sigma = \sigma^2 \mathrm{I}$
- Decision rule $= (\mu_0 - \mu_1)^T \Sigma^{-1} x + constants$

$$(\mu_0 - \mu_1)^T x \underset{>}{\overset{<}{}} C$$



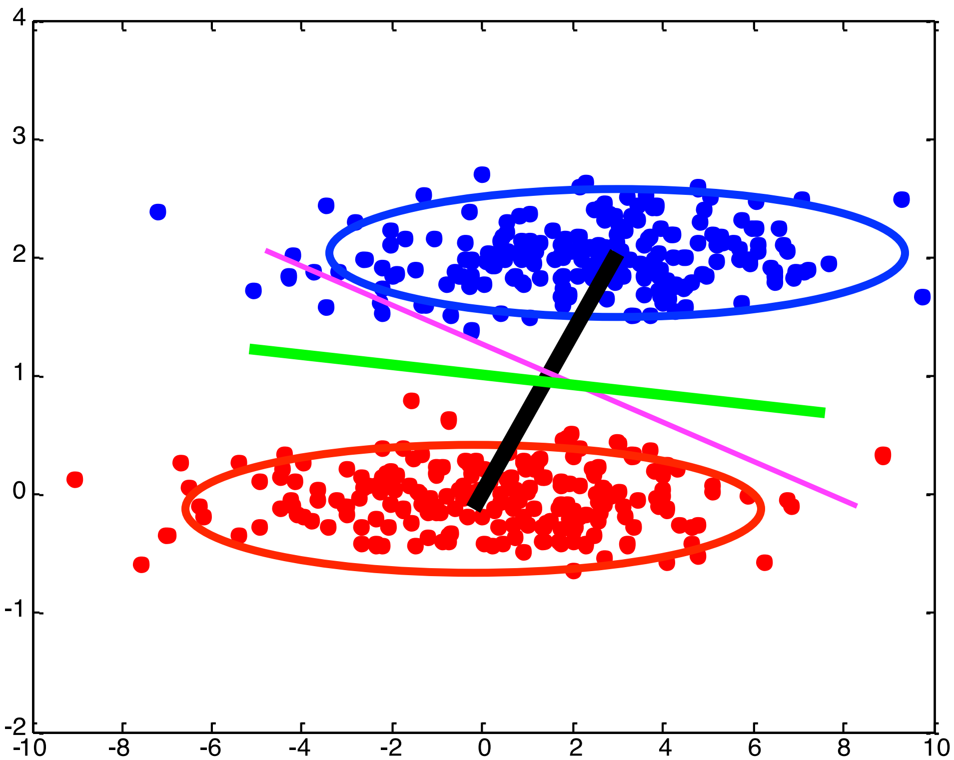
$$(\mu_0 - \mu_1)$$

$$C = .5(\mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1) - \log \frac{p(y=0)}{p(y=1)}$$

Non-spherical Gaussian distributions

- Equal covariances => still linear decision rule
  - May be “modulated” by variance direction
  - Scales; rotates (if correlated)

Ex:
Variance
[3 0]
[0 .25]



Class posterior probabilities

- Useful to also know class *probabilities*
- Some notation
  - p(y=0) , p(y=1) – class *prior* probabilities
    - How likely is each class in general?
  - p(x | y=c) – class conditional probabilities
    - How likely are observations "x" in that class?
  - p(y=c | x) – class posterior probability
    - How likely is class c *given* an observation x?

Class posterior probabilities

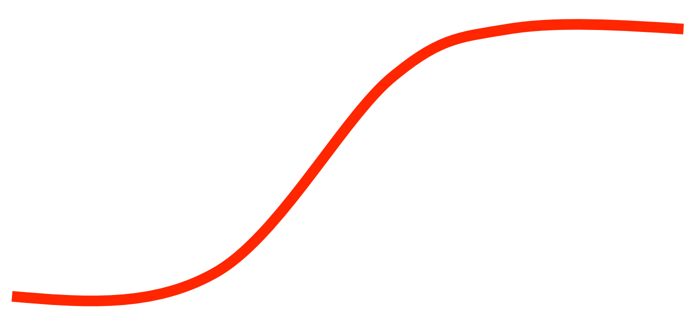
- Useful to also know class *probabilities*
- Some notation
  - p(y=0) , p(y=1) – class *prior* probabilities
    - How likely is each class in general?
  - p(x | y=c) – class conditional probabilities
    - How likely are observations "x" in that class?
  - p(y=c | x) – class posterior probability
    - How likely is class c *given* an observation x?

- We can compute posterior using Bayes' rule
  - p(y=c | x) = p(x|y=c) p(y=c) / p(x)
- Compute p(x) using sum rule / law of total prob.
  - p(x) = p(x|y=0) p(y=0) + p(x|y=1)p(y=1)

Class posterior probabilities

- Consider comparing two classes
  - p(x | y=0) * p(y=0)   vs   p(x | y=1) * p(y=1)
  - Write probability of each class as
  - p(y=0 | x) = p(y=0, x) / p(x)
  -                 = p(y=0, x) / (p(y=0,x) + p(y=1,x))
  -     = 1 / (1 + exp(-a))   (**)

  - a = log [p(x|y=0) p(y=0) / p(x|y=1) p(y=1)]
  - (**) called the logistic function, or logistic sigmoid.



Gaussian models

- Return to Gaussian models with equal covariances

$$\mathcal{N}(\underline{x}\ ;\ \underline{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2}} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}) \right\}$$

$$0 \begin{matrix} < \\ > \end{matrix} \ \log \frac{p(x|y=0)}{p(x|y=1)} \frac{p(y=0)}{p(y=1)} = \quad (\mu_0 - \mu_1)^T \Sigma^{-1} x + constants$$

(**)

Now we also know that the probability of each class is given by:
p(y=0 | x) = Logistic(**) = Logistic(a^T x + b)

We'll see this form again soon…

Summary

- Axioms of probability
  - Help us reason explicitly about uncertainty
- Random variables
- Discrete variables; probability mass functions
  - Positive values, sum to one
  - Bernoulli, Discrete, etc.
- Joint distributions
  - Law of total probability
  - Chain rule of conditional probability
- Continuous variables; probability density functions
  - Gaussian