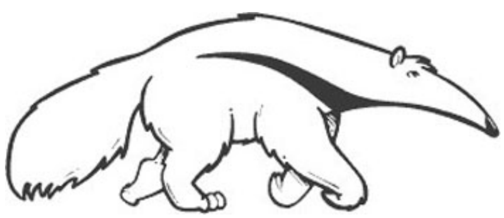
Machine Learning and Data Mining

Bayes Classifiers

Prof. Alexander Ihler



BREN:ICS INFORMATION AND COMPUTER SCIENCES

UNIVERSITY of CALIFORNIA IRVINE

A basic classifier

- Training data D={$x^{(i)}$,$y^{(i)}$}, Classifier f(x ; D)
  - Discrete feature vector x
  - f(x ; D) is a contingency table
- Ex: credit rating prediction (bad/good)
  - X_1 = income (low/med/high)
  - How can we make the most # of correct predictions?

Features	# bad	# good
X=0	42	15
X=1	338	287
X=2	3	5

A basic classifier

- Training data $D=\{x^{(i)},y^{(i)}\}$, Classifier $f(x ; D)$
  - Discrete feature vector x
  - f(x ; D) is a contingency table
- Ex: credit rating prediction (bad/good)
  - X_1 = income (low/med/high)
  - How can we make the most # of correct predictions?
  - Predict more likely outcome
    for each possible observation

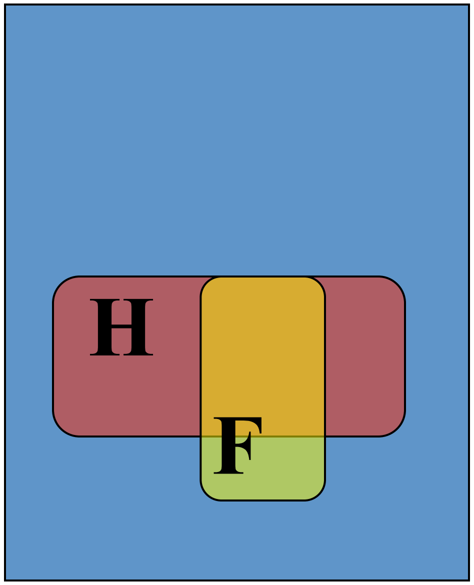
Features	# bad	# good
X=0	42	15
X=1	338	287
X=2	3	5

A basic classifier

- Training data $D=\{x^{(i)},y^{(i)}\}$, Classifier $f(x ; D)$
  - Discrete feature vector x
  - f(x ; D) is a contingency table
- Ex: credit rating prediction (bad/good)
  - X_1 = income (low/med/high)
  - How can we make the most # of correct predictions?
  - Predict more likely outcome
    for each possible observation
  - Can normalize into probability:
    p(y=good | X=c)
  - How to generalize?

Features	# bad	# good
X=0	.7368	.2632
X=1	.5408	.4592
X=2	.3750	.6250

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?

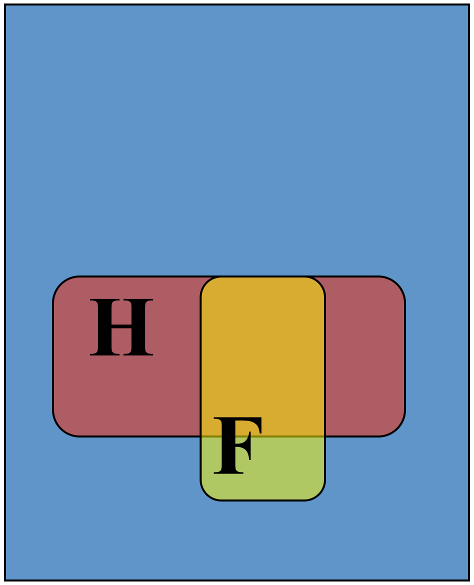
Example from Andrew Moore's slides

Bayes rule

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

- $P(H \& F) = ?$

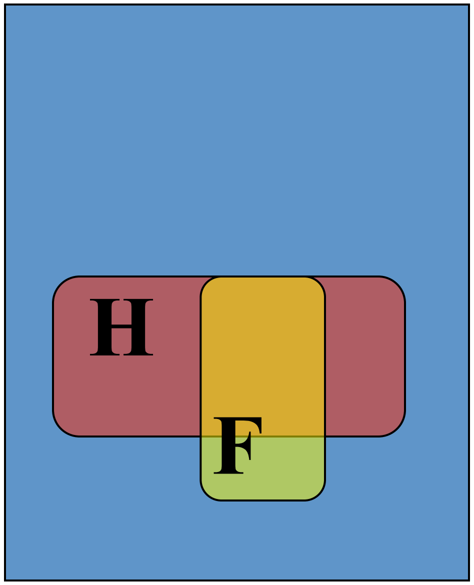
- $P(F|H) = ?$



Example from Andrew Moore's slides

Bayes rule

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

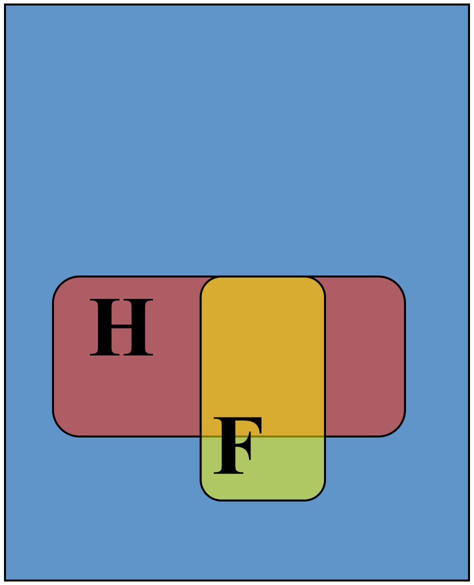


- $P(H \& F) = p(F)\, p(H|F)$
  $= (1/2) * (1/40) = 1/80$
- $P(F|H) = ?$

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(y)
  - E.g., fraction of applicants that have good credit
- Distribution of features given the class, p(x | y=c)
  - How likely are we to see "x" in users with good credit?

- Joint distribution $$p(y|x)p(x) = p(x,y) = p(x|y)p(y)$$

- Bayes Rule: $$\Rightarrow \quad p(y|x) = p(x|y)p(y)/p(x)$$

$$= \frac{p(x|y)p(y)}{\sum_c p(x|y=c)p(y=c)}$$

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

- For a discrete x, this recalculates the same table…

Features	# bad	# good
X=0	42	15
X=1	338	287
X=2	3	5

$\Rightarrow$

p(x | y=0)	p(x | y=1)
42 / 383	15 / 307
338 / 383	287 / 307
3 / 383	5 / 307

$\Rightarrow$

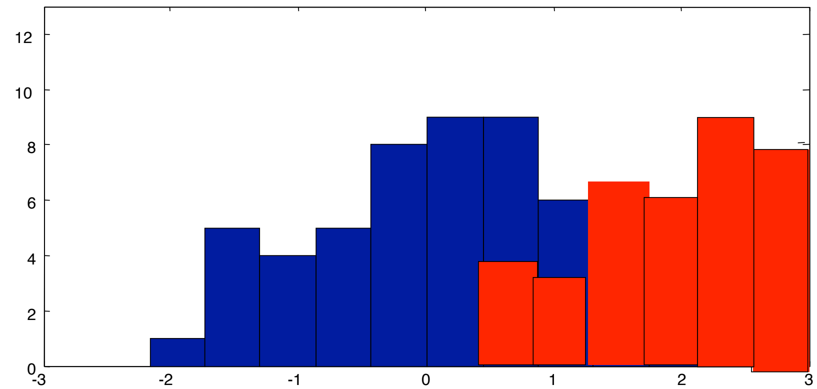
p(y=0|x)	p(y=1|x)
.7368	.2632
.5408	.4592
.3750	.6250

p(y)	**383/690**	**307/690**

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

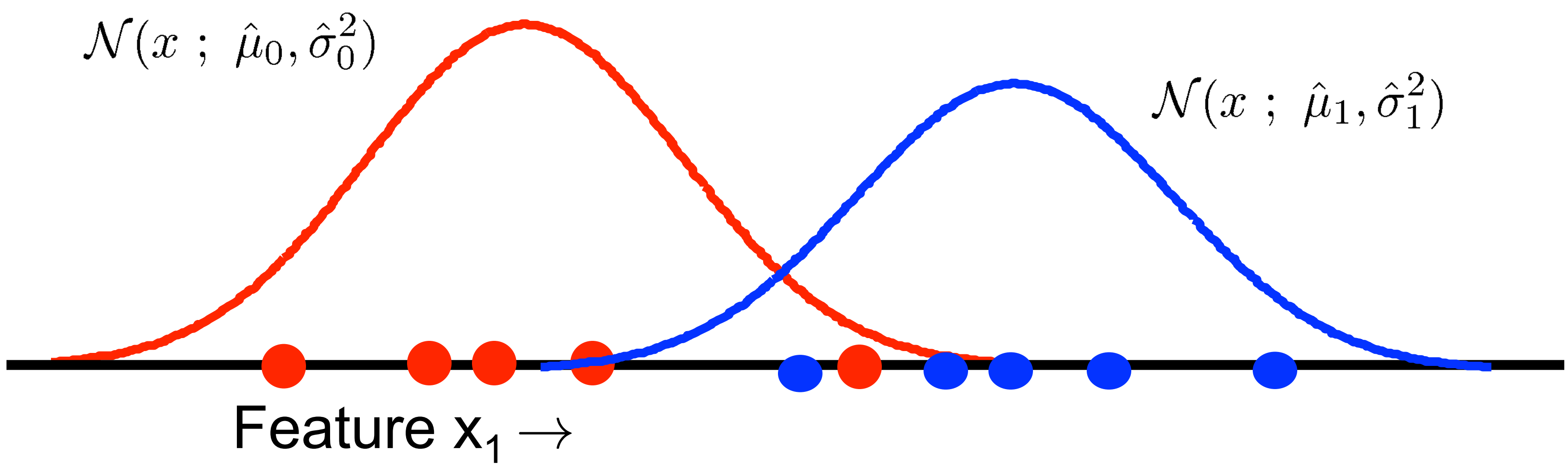
- For continuous x, can use any density estimate we like
  - Histogram
  - Gaussian
  - …



Gaussian models

- Estimate parameters of the Gaussians from the data

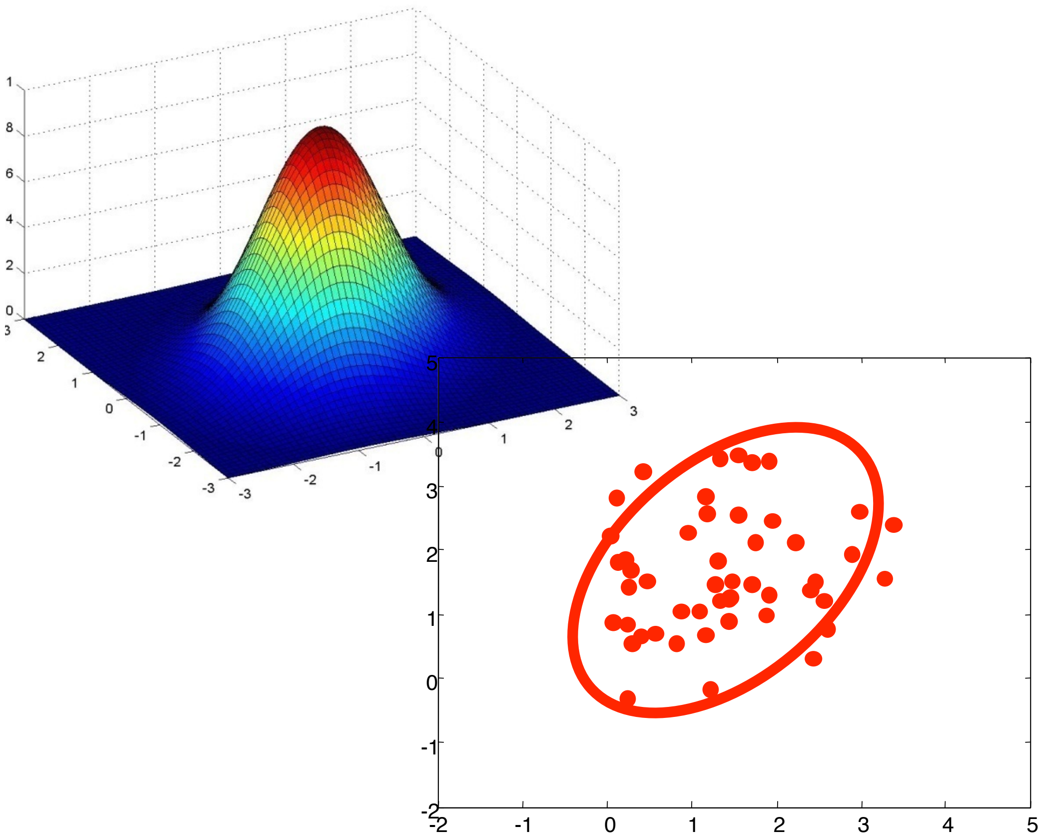
$$\alpha = \frac{m_1}{m} = \hat{p}(y = c_1) \qquad \hat{\mu} = \frac{1}{m}\sum_j x^{(j)} \qquad \hat{\sigma}^2 = \frac{1}{m}\sum_j (x^{(j)} - \mu)^2$$



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:

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

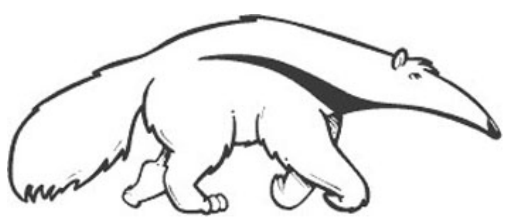
$$\hat{\Sigma} = \frac{1}{m} \sum_j (\underline{x}^{(j)} - \underline{\hat{\mu}})^T (\underline{x}^{(j)} - \underline{\hat{\mu}})$$

+

Machine Learning and Data Mining

Bayes Classifiers: Naïve Bayes

Prof. Alexander Ihler



BREN:ICS INFORMATION AND COMPUTER SCIENCES

UNIVERSITY of CALIFORNIA IRVINE

Bayes classifiers

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

- For a discrete x, can represent as a contingency table…
  – What about if we have more discrete features?

Features	# bad	# good
X=0	42	15
X=1	338	287
X=2	3	5

$\Rightarrow$

p(x | y=0)	p(x | y=1)
42 / 383	15 / 307
338 / 383	287 / 307
3 / 383	5 / 307

$\Rightarrow$

p(y=0|x)	p(y=1|x)
.7368	.2632
.5408	.4592
.3750	.6250

p(y)	**383/690**	**307/690**

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(A,B,C)
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

A	B	C	p(A,B,C)
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?
- Overfitting!

Overfitting and density estimation

A	B	C	p(A,B,C)
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:
- $p(a,b) = p(a)\,p(b)$

- $p(x_1, x_2, \ldots x_N) = p(x_1)\,p(x_2) \ldots p(x_N)$
- Only need to estimate each individually

A	p(A)
0	.4
1	.6

B	p(B)
0	.7
1	.3

C	p(C)
0	.1
1	.9

$\Rightarrow$

A	B	C	p(A,B,C)
0	0	0	.4 * .7 * .1
0	0	1	.4 * .7 * .3
0	1	0	.4 * .3 * .1
0	1	1	…
1	0	0	
1	0	1	
1	1	0	
1	1	1	

Naïve Bayes models

- Variable y to predict, e.g. "auto accident in next year?"
- We have *many* co-observed vars **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:
  - $p(y|\mathbf{x}) = p(\mathbf{x}|y)\, p(y) / p(\mathbf{x})$; $p(\mathbf{x}|y) = \prod_\iota p(x_i|y)$
  - Covariates are independent given "cause"
- Note: 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

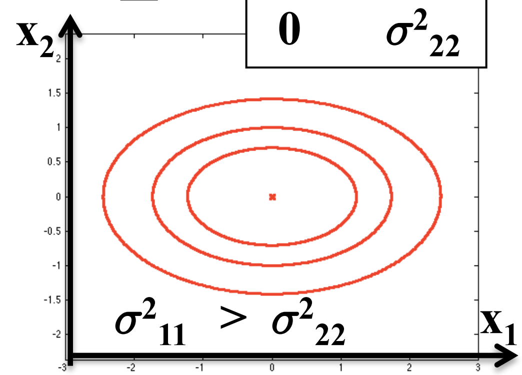
$$p(x_1) = \frac{1}{Z} \exp\left\{-\frac{1}{2\sigma_1^2}(x_1-\mu_1)^2\right\} \qquad p(x_2) = \frac{1}{Z_2} \exp\left\{-\frac{1}{2\sigma_2^2}(x_2-\mu_2)^2\right\}$$

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

$$\underline{\mu} = [\mu_1 \;\; \mu_2]$$

$$\Sigma = \mathrm{diag}(\sigma_1^2 \;,\;\; \sigma_2^2)$$

$$\Sigma = \boxed{\begin{matrix} \sigma^2_{11} & 0 \\ 0 & \sigma^2_{22} \end{matrix}}$$



Again, reduces the number of parameters of the model:
Bayes: $m^2/2$ Naïve Bayes: m

You should know…

- Bayes rule; $p(y \mid x)$
- Bayes classifiers
  - Learn $p(x \mid y=C)$, $p(y=C)$
- Naïve Bayes classifiers
  - Assume features are independent given class:

    $$p(x \mid y=C) = p(x_1 \mid y=C)\, p(x_2 \mid y=C) \ldots$$

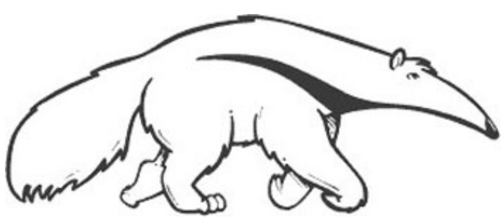
- Maximum likelihood (empirical) estimators for
  - Discrete variables
  - Gaussian variables
  - Overfitting; simplifying assumptions or regularization

+

Machine Learning and Data Mining

Bayes Classifiers: Measuring Error

Prof. Alexander Ihler



BREN:ICS
INFORMATION AND COMPUTER SCIENCES

UNIVERSITY of CALIFORNIA IRVINE

A Bayes classifier

- Given training data, compute p(y=c| x) and choose largest
- What's the error rate of this method?

Features	# bad	# good
X=0	42	15
X=1	338	287
X=2	3	5

A Bayes classifier

- Given training data, compute p(y=c| x) and choose largest
- What's the error rate of this method?

Features	# bad	# good
X=0	42	15
X=1	338	287
X=2	3	5

Gets these examples wrong:

Pr[error] = (15 + 287 + 3) / (690)

(empirically on training data:
better to use test data)

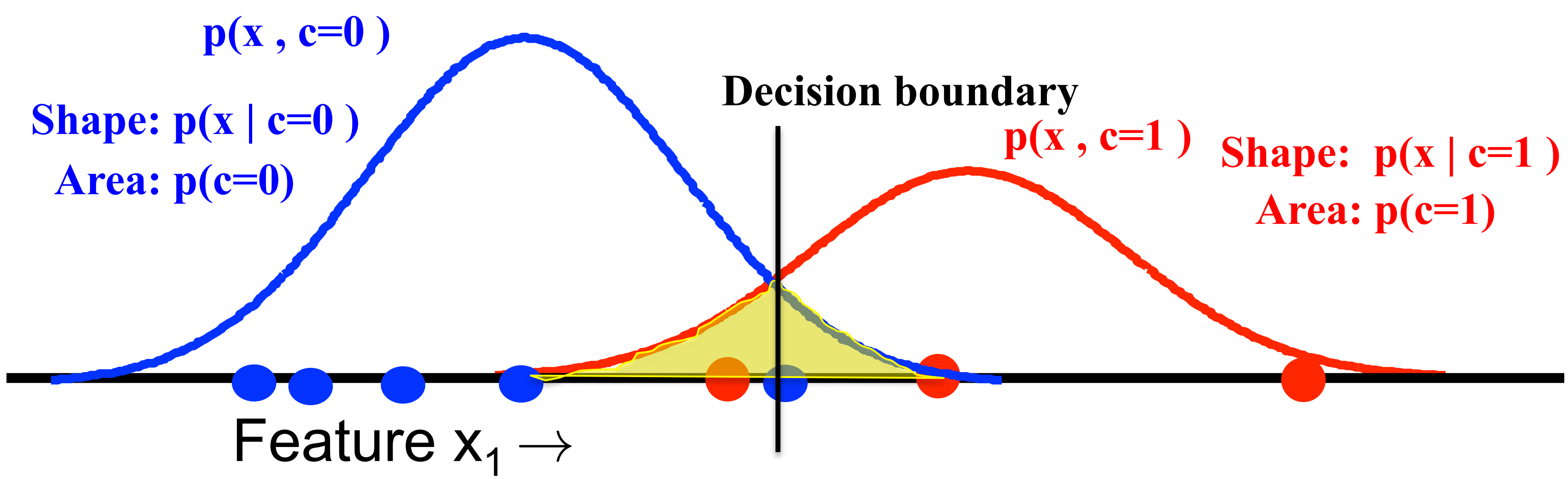
A Bayes classifier

- Similar form & computation for continuous x

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

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

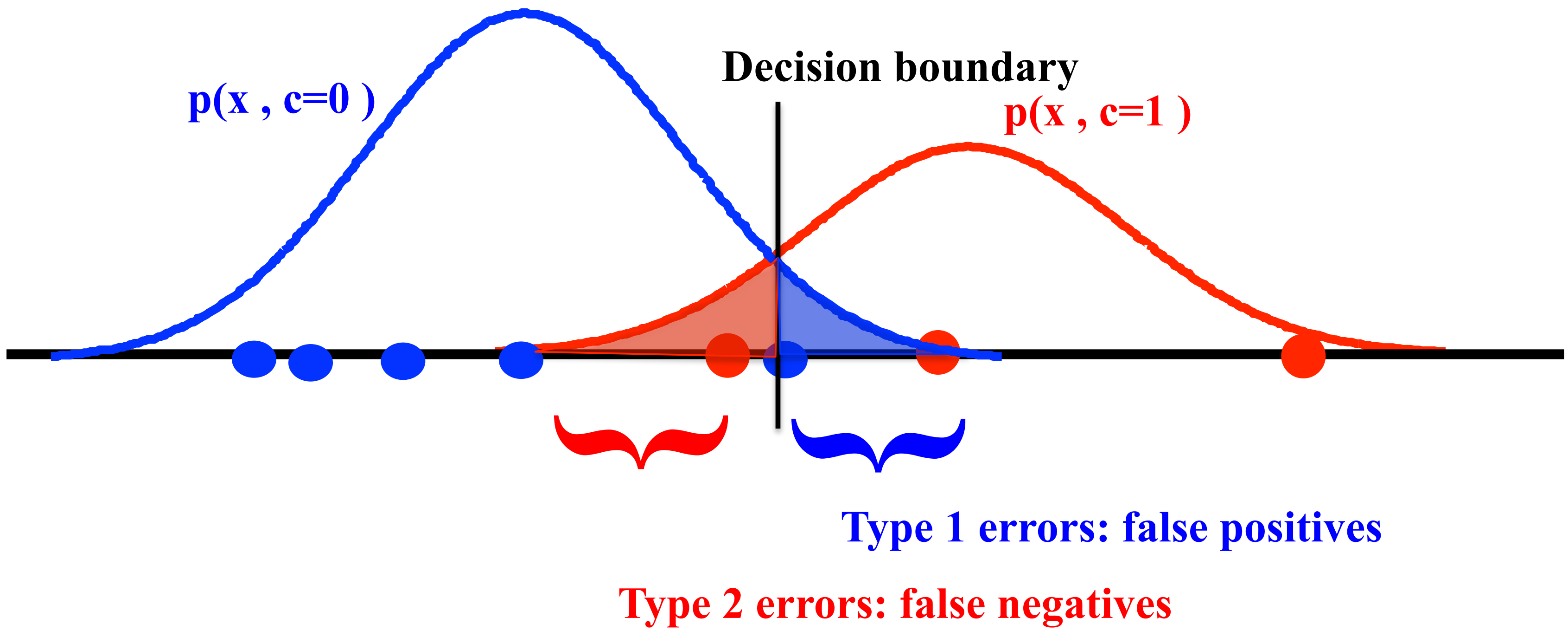
"log likelihood ratio"



A Bayes classifier

- Not all errors are created equally…
- Risk associated with each outcome?

$$\gamma \quad \begin{matrix} < \\ > \end{matrix} \quad \log \frac{p(x|y=1)}{p(x|y=0)}$$



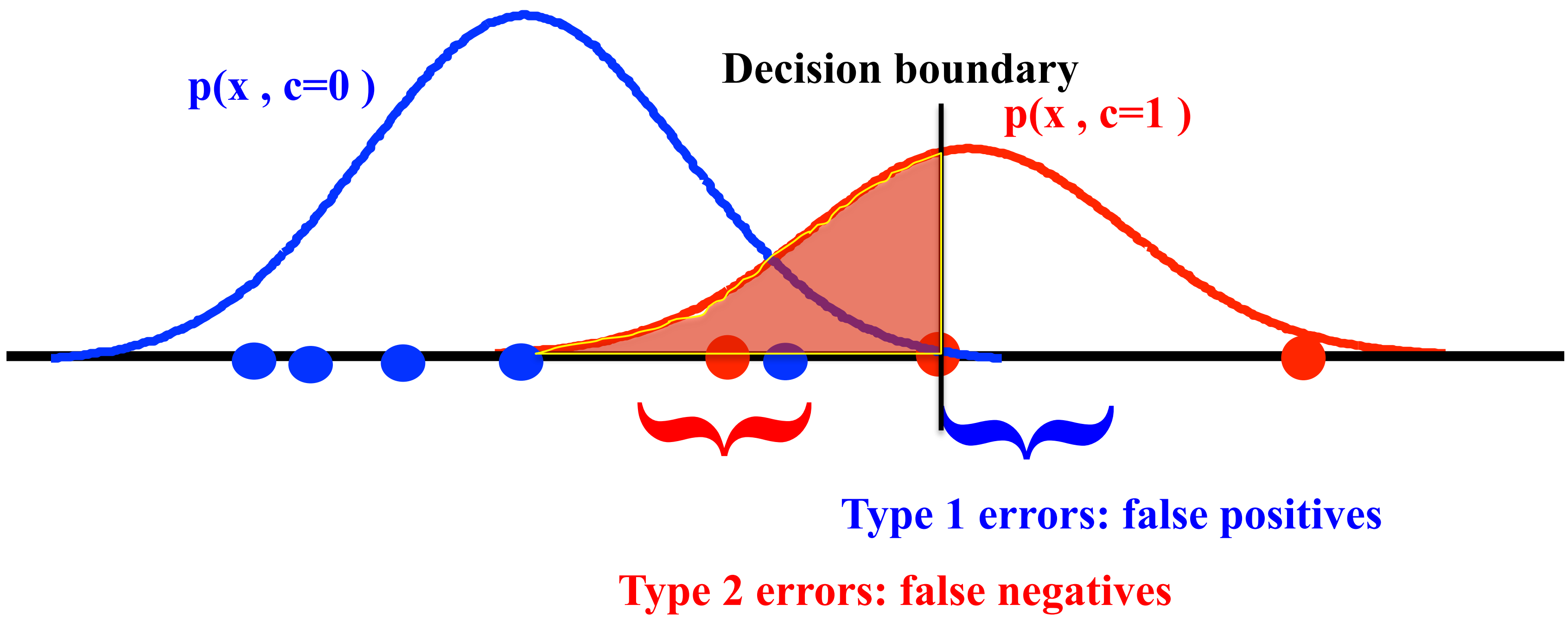
False positive rate: (# y=0, ŷ=1) / (#y=0)

False negative rate: (# y=1, ŷ=0) / (#y=1)

A Bayes classifier

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

- Increase gamma: prefer class 0
- Spam detection



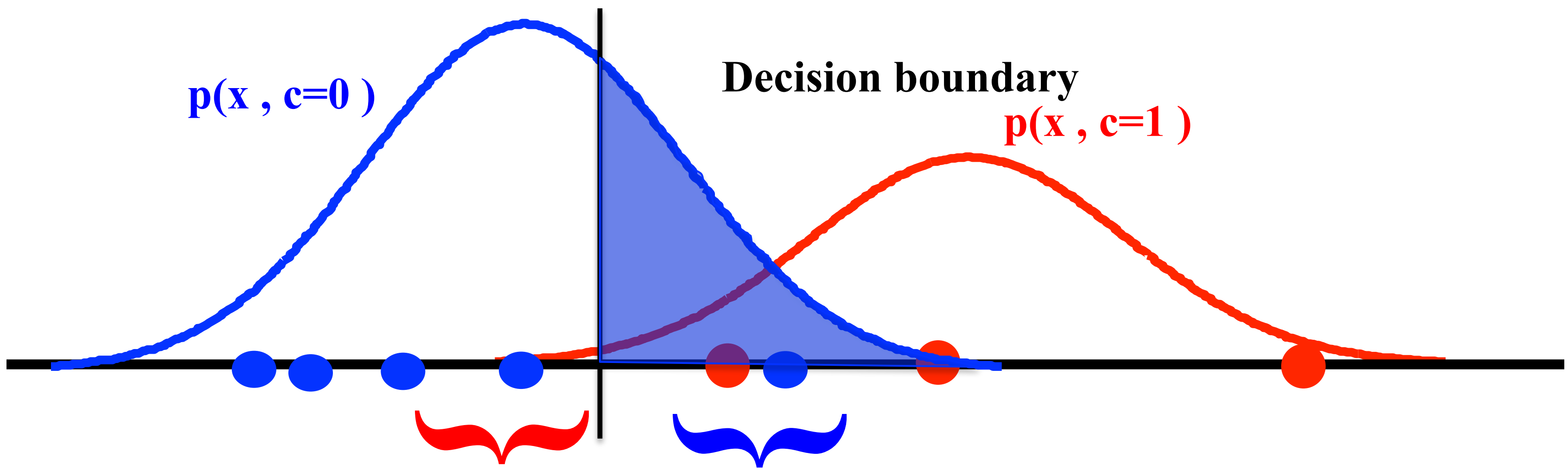
False positive rate: (# y=0, ŷ=1) / (#y=0)

False negative rate: (# y=1, ŷ=0) / (#y=1)

A Bayes classifier

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

- Decrease gamma: prefer class 1
- Cancer detection



Type 1 errors: false positives

Type 2 errors: false negatives

False positive rate: (# y=0, ŷ=1) / (#y=0)

False negative rate: (# y=1, ŷ=0) / (#y=1)

Measuring errors

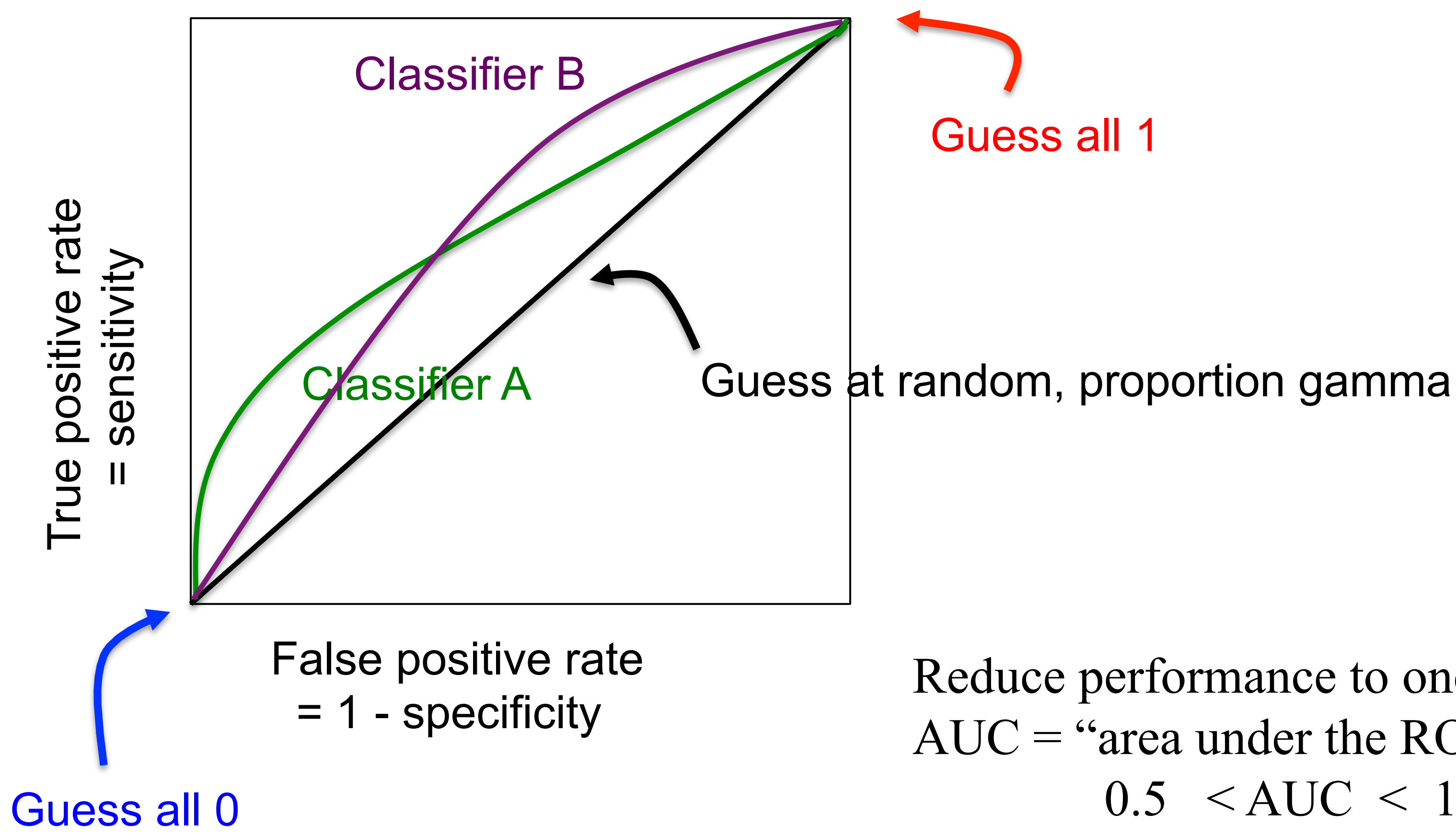
- Confusion matrix
- Can extend to more classes

	Predict 0	Predict 1
Y=0	380	5
Y=1	338	3

- True positive rate: #(y=1 , ŷ=1) / #(y=1) -- "sensitivity"
- False negative rate: #(y=1 , ŷ=0) / #(y=1)
- False positive rate: #(y=0 , ŷ=1) / #(y=0)
- True negative rate: #(y=0 , ŷ=0) / #(y=0) -- "specificity"

ROC Curves

- Characterize performance over various gamma?



Reduce performance to one number?
AUC = "area under the ROC curve"
$0.5 < \text{AUC} < 1$