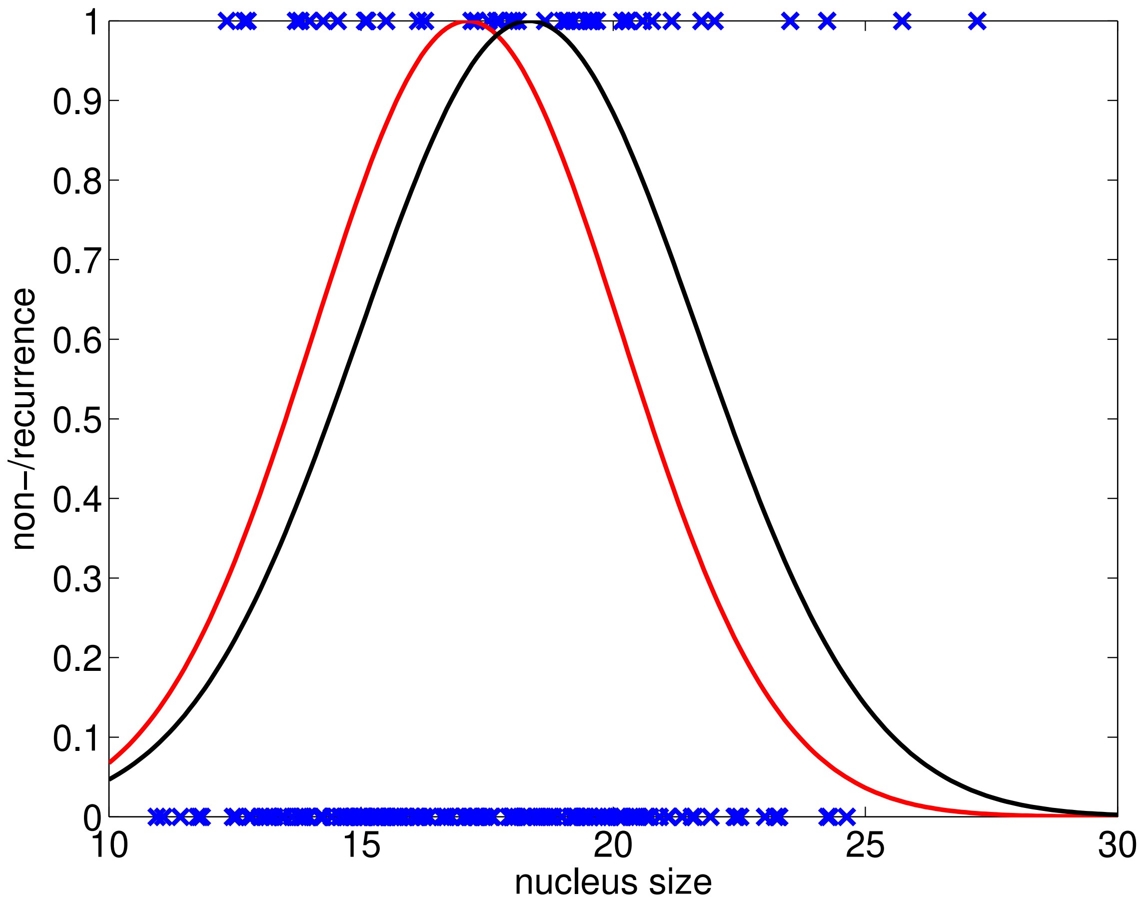
COMP 652: Machine Learning

Lecture 8: More generative learning

Recalling generating learning for classification



- Model the input distributions for each class
- Predict class of new vector $\mathbf{x}$ based on these distributions

More formally

- Suppose we have data of the form $\langle \mathbf{x}, y \rangle$, where $\mathbf{x}$ is a feature vector and $y \in \mathcal{Y}$ is a class label.
- Generative learning algorithms model two things:
  - $P(y)$ – the *prior probability* of each possible class $y \in \mathcal{Y}$
  - $P(\mathbf{x}|y)$ – the *class conditional input distributions*
- We can then estimate the probability that a new input $\mathbf{x}$ is of class y be a simple application of Bayes's rule:

$$P(y|\mathbf{x}) = \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})} = \frac{P(\mathbf{x}|y)P(y)}{\sum_{y' \in \mathcal{Y}} P(\mathbf{x}|y')P(y')}$$

Details of Wisconsin example

- $P(y=0) = \frac{148}{194} = 0.7629$
- $P(y=1) = \frac{46}{194} = 0.2371$
- Among non-recurrers $(y=0)$, mean nucleus size is $\mu_0 = 17.1137$ and standard deviation is $\sigma_0 = 3.0637$.
  We approximate $P(x|y=0)$ as Normal(μ_0,σ_0).
- Among recurrers $(y=1)$, mean nucleus size is $\mu_1 = 18.3309$ and standard deviation is $\sigma_1 = 3.3656$.
  We approximate $P(x|y=1)$ as Normal(μ_1,σ_1).

Details of Wisconsin example (II)

- For a new x, we predict

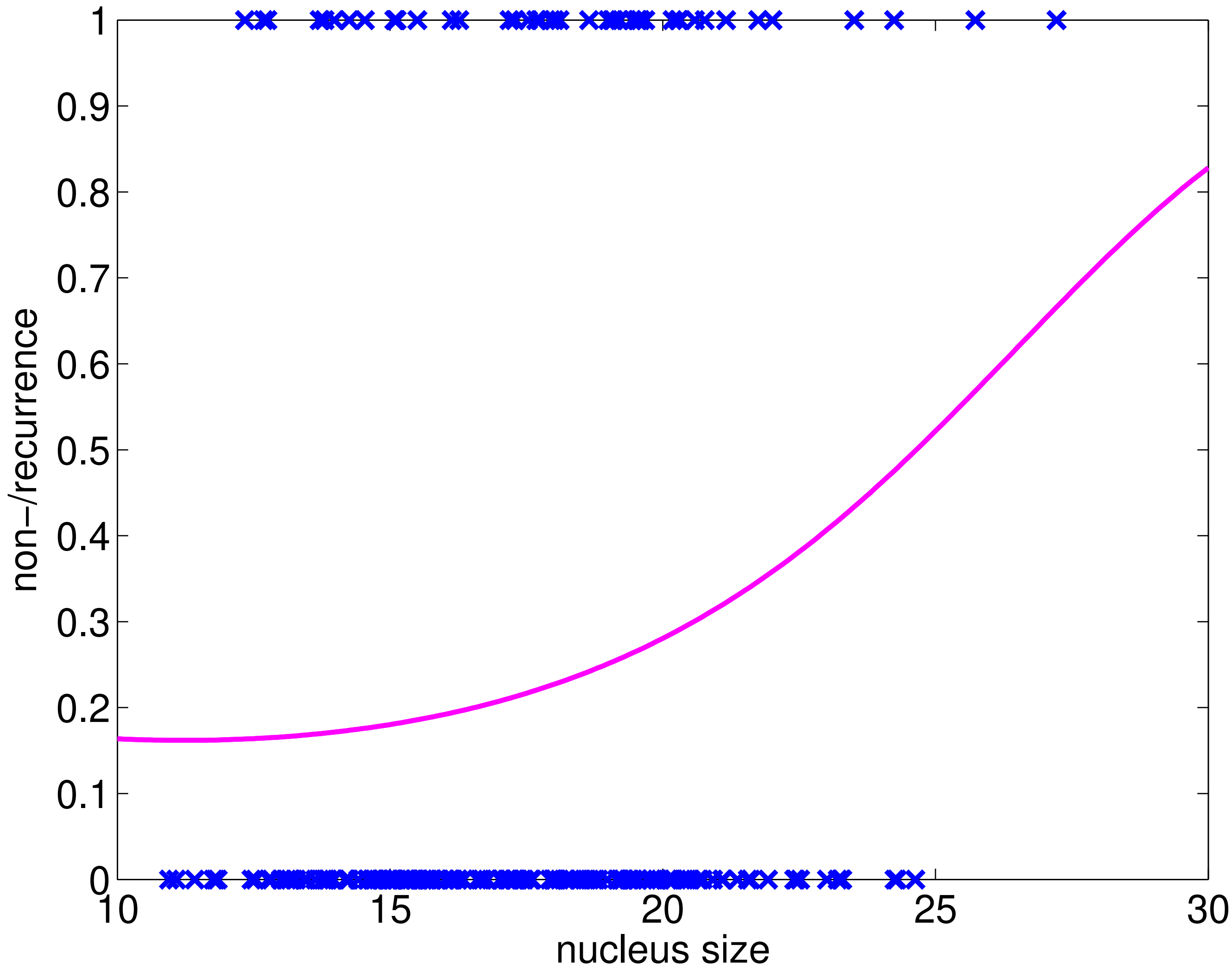
$$P(y=1|x) = \frac{P(x|y=1)P(y=1)}{P(x)}$$

$$= \frac{P(x|y=1)P(y=1)}{P(x|y=0)P(y=0)+P(x|y=1)P(y=1)}$$

$$= \frac{\frac{1}{\sqrt{2\pi}\sigma_1}\exp(-\frac{(x-\mu_1)^2}{2\sigma_1^2})P(y=1)}{\frac{1}{\sqrt{2\pi}\sigma_0}\exp(-\frac{(x-\mu_0)^2}{2\sigma_0^2})P(y=0)+\frac{1}{\sqrt{2\pi}\sigma_1}\exp(-\frac{(x-\mu_1)^2}{2\sigma_1^2})P(y=1)}$$

- $P(y=0|x)$ can be computed similarly, or simply as $1-P(y=0|x)$.

Visualization of $P(y = 1|x)$



Density estimation

- The basic problem: given a set of data $\{\mathbf{x}_i\}$, estimate $P(\mathbf{x})$.
- If the data comes from a discrete set, then we are estimating a discrete distribution.
- When the data is real-valued, we (typically) estimate a continuous density function.
- We call it *parametric density estimation* when we assume that $P(\mathbf{x})$ follows a typical, parameterized density function—such as Gaussian, Exponential, Bernoulli, Binomial, etc.
- In later lectures, we will consider semi-parametric and non-parametric approaches.

Univariate Gaussian density estimation

- Suppose we have univariate data: $x_1, x_2, \ldots, x_m$.
- Suppose we assume that the data are i.i.d. observations from a Gaussian distribution with unknown mean μ and unknown standard deviation σ.
- Suppose we estimate μ and σ by finding the values that maximize the likelihood of the data:

$$P(x_1, x_2, \ldots, x_m | \mu, \sigma) = \Pi_{i=1}^{m} \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right)$$

Derivation of maximum likelihood μ and σ

$$
\begin{aligned}
& \arg\max_{\mu,\sigma} \; P(x_1, x_2, \ldots, x_m | \mu, \sigma) \\
= \; & \arg\max_{\mu,\sigma} \; \Pi_{i=1}^{m} \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right) \\
\equiv \; & \arg\max_{\mu,\sigma} \; \log \left[\Pi_{i=1}^{m} \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right) \right] \\
= \; & \arg\max_{\mu,\sigma} \; \Sigma_{i=1}^{m} \left[-\log(\sqrt{2\pi}\sigma) - \frac{(x_i-\mu)^2}{2\sigma^2} \right] \\
= \; & \arg\max_{\mu,\sigma} \; -m\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \left(\Sigma_{i=1}^{m} (x_i-\mu)^2 \right)
\end{aligned}
$$

Derivation of maximum likelihood μ and σ (II)

Set the partial derivative with respect to μ equal to zero:

$$\begin{aligned}
\frac{1}{\sigma^2}\left(\Sigma_{i=1}^m (x_i - \mu)\right) &= 0 \\
\Sigma_{i=1}^m (x_i - \mu) &= 0 \\
\Sigma_{i=1}^m x_i - m\mu &= 0 \\
\mu &= \frac{\Sigma_{i=1}^m x_i}{m}
\end{aligned}$$

The maximum likelihood choice for μ is just the sample average.

Derivation of maximum likelihood μ and σ (III)

Set the partial derivative with respect to σ equal to zero:

$$\begin{aligned}
\frac{-m}{\sigma} + \frac{1}{\sigma^3}\left(\Sigma_{i=1}^{m}(x_i-\mu)^2\right) &= 0 \\
-m + \frac{1}{\sigma^2}\left(\Sigma_{i=1}^{m}(x_i-\mu)^2\right) &= 0 \\
\frac{1}{\sigma^2} &= \frac{m}{\Sigma_{i=1}^{m}(x_i-\mu)^2} \\
\sigma^2 &= \frac{\Sigma_{i=1}^{m}(x_i-\mu)^2}{m} \\
\sigma &= \sqrt{\frac{\Sigma_{i=1}^{m}(x_i-\mu)^2}{m}}
\end{aligned}$$

The maximum likelihood choice for σ^2 is sample variance, and for σ is the sample standard deviation.

Recall multivariate Gaussian distribution

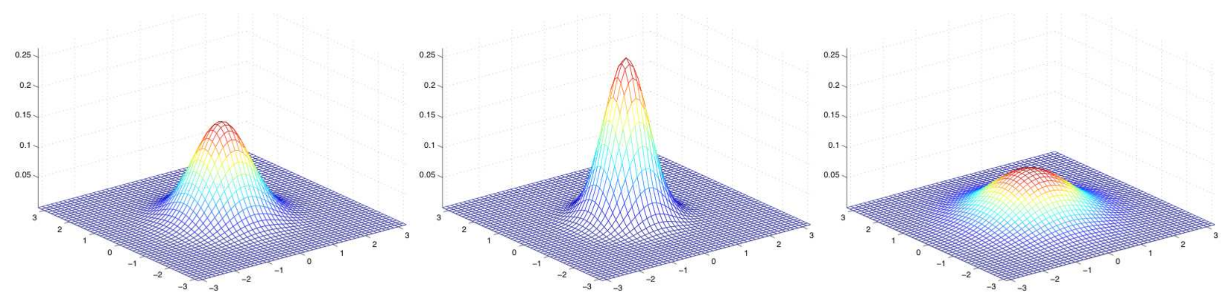
- If $\mathbf{x}$ is a vector of length n, then a multivariate Gaussian distribution on $\mathbf{x}$ is defined by
  - A mean vector $\mu \in \Re^n$.
  - An $n \times n$ symmetric positive definite covariance matrix Σ. $\Sigma_{i,j}$ gives the covariance of x_i and x_j: $E((x_i - \mu_i)(x_j - \mu_j))$.
- The density function is given by

$$P(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right)$$

Examples (I)

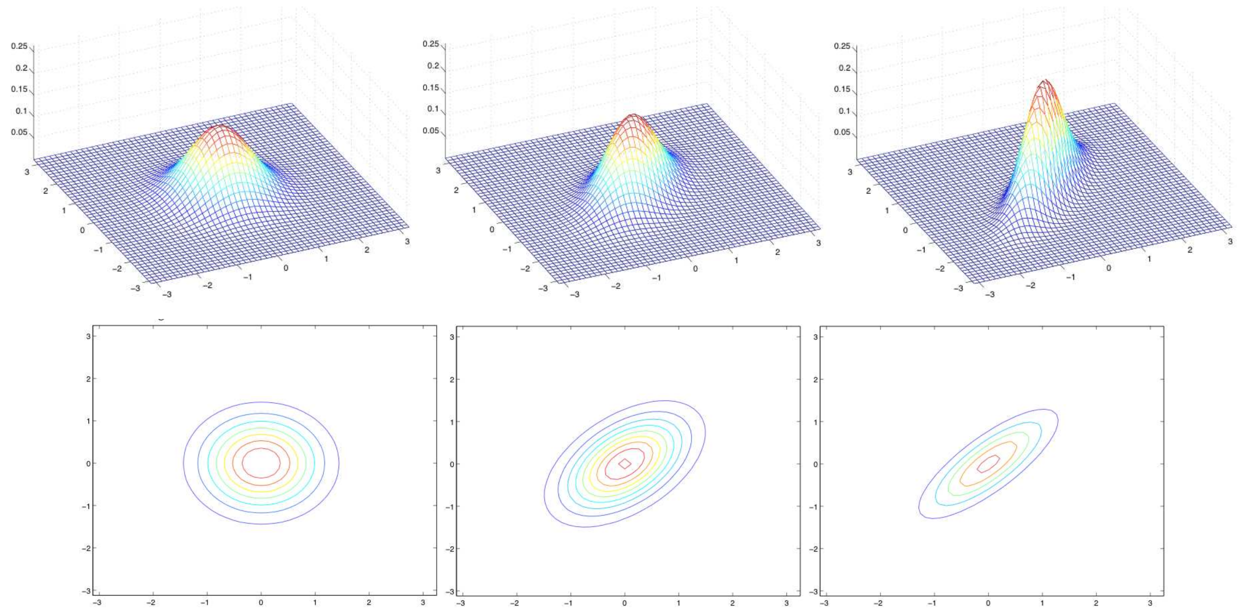
(All with $\mu = 0$.)

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$



Examples (II)

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$



Maximum likelihood parameter estimates

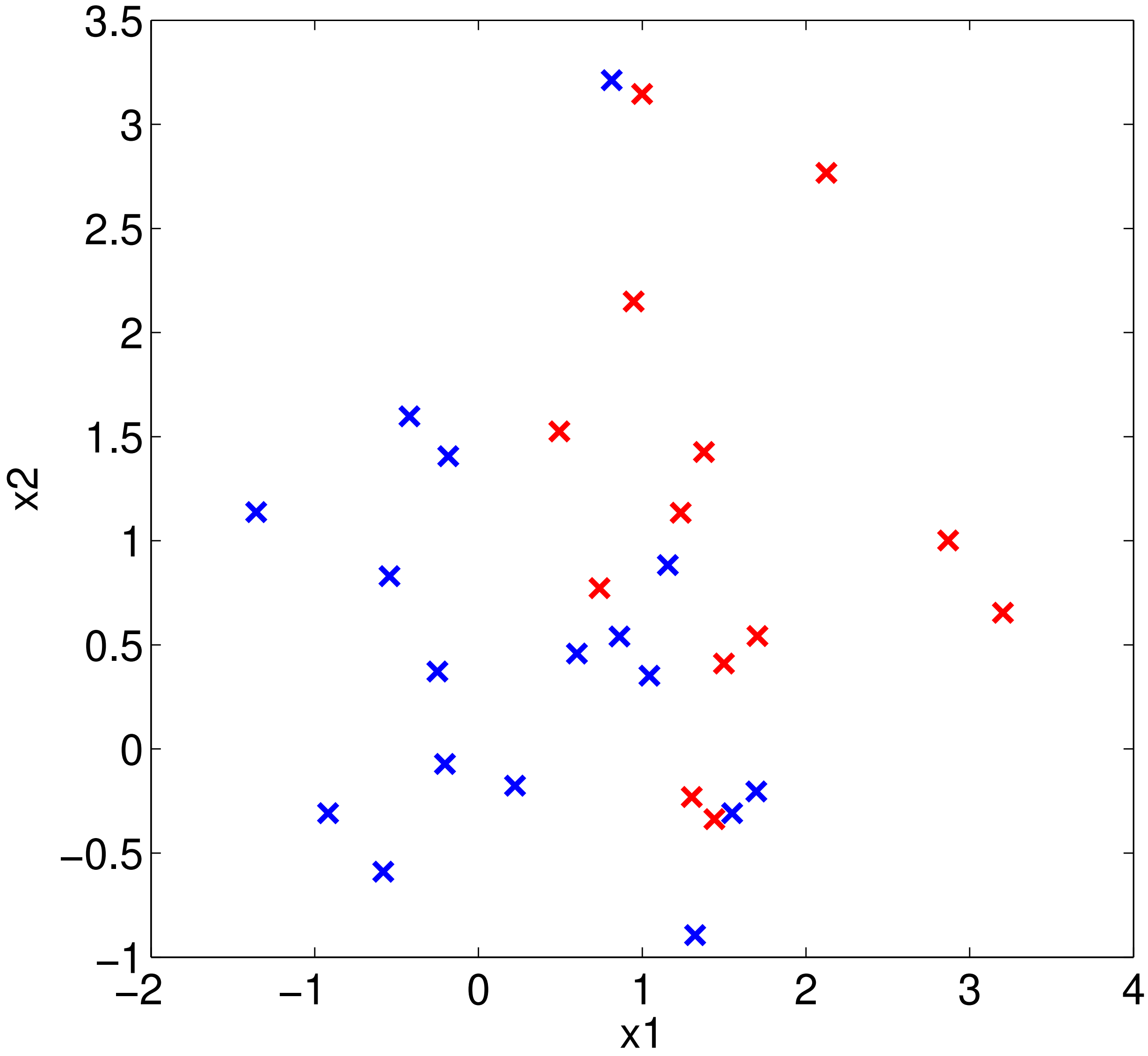
- Suppose we have a data set $\mathbf{x_1}$, $\mathbf{x_2}$, ..., $\mathbf{x_m}$ of vectors in $\Re^n$.
- Suppose we assume a multivariate Gaussian distribution for the $\mathbf{x_i}$.
- The maximum likelihood estimates for μ and Σ can be derived similarly to the univariate case, and are:
  - $\mu = \frac{1}{m}\Sigma_{i=1}^{m}\mathbf{x_i}$ (the sample mean)
  - $\Sigma_{j,k} = \frac{1}{m}\Sigma_{i=1}^{m}(x_{i,j} - \mu_j)(x_{i,k} - \mu_k)$ (the sample covariance)

    Equivalently, $\Sigma = \frac{1}{m}\Sigma_{i=1}^{m}(\mathbf{x_i} - \mu)(\mathbf{x_i} - \mu)^T$.
- This involves fitting $\frac{n^2}{2} + \frac{n}{2}$ parameters. If we don't have enough data, we may make a simplifying assumtpion, such as:
  - All covariances are zero, so that Σ is diagonal, and we estimate only $2n$ parameters.
  - All covariances are zero, and all variances are σ^2, so that $\Sigma = \sigma^2 I$, and we estimate only $n + 1$ parameters.

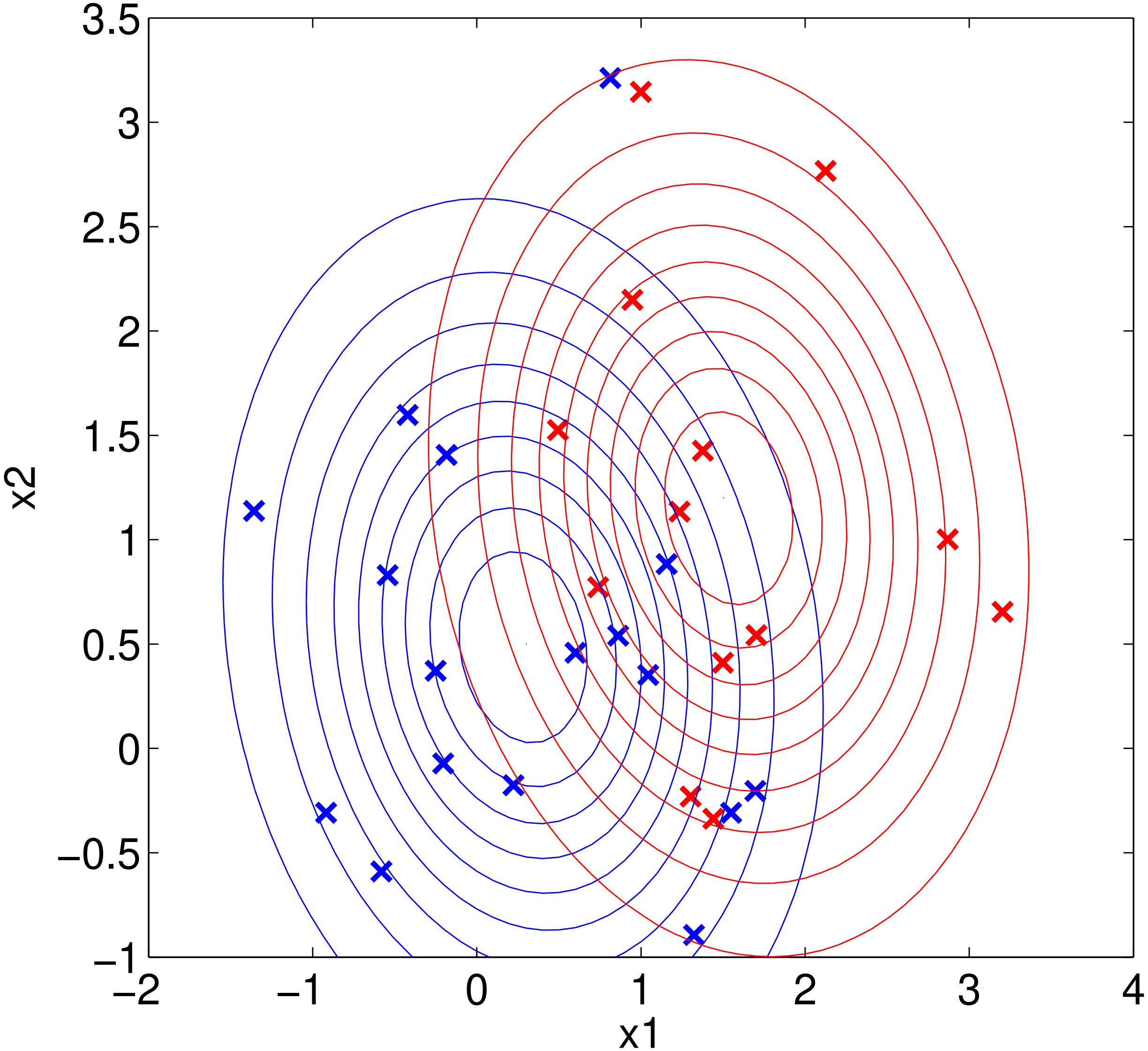
Gaussian discriminant analysis

- *Gaussian discriminant analysis* (GDA) is a generative learning approach in which the input distributions are assume to be multivariate Gaussian. (Our approach on the Wisconsin example at the start of this lecture was an instance of GDA.)
- In *linear discriminant analysis* (LDA), we assume that:
  - Each class $y \in \mathcal{Y}$ is associated with it's own mean vector μ_y.
  - All classes have the same covariance matrix Σ, so that $P(\mathbf{x}|y) \sim N(\mu_u, \Sigma)$
  - These parameters are determined by maximum likelihood:
    - $\mu_y = (\sum_{i:y_i=y} \mathbf{x_i})/|\{i : y_i = y\}|$
    - $\Sigma = \frac{1}{m} \sum_{i=1}^{m} (\mathbf{x_i} - \mu_{y_i})(\mathbf{x_i} - \mu_{y_i})^T$

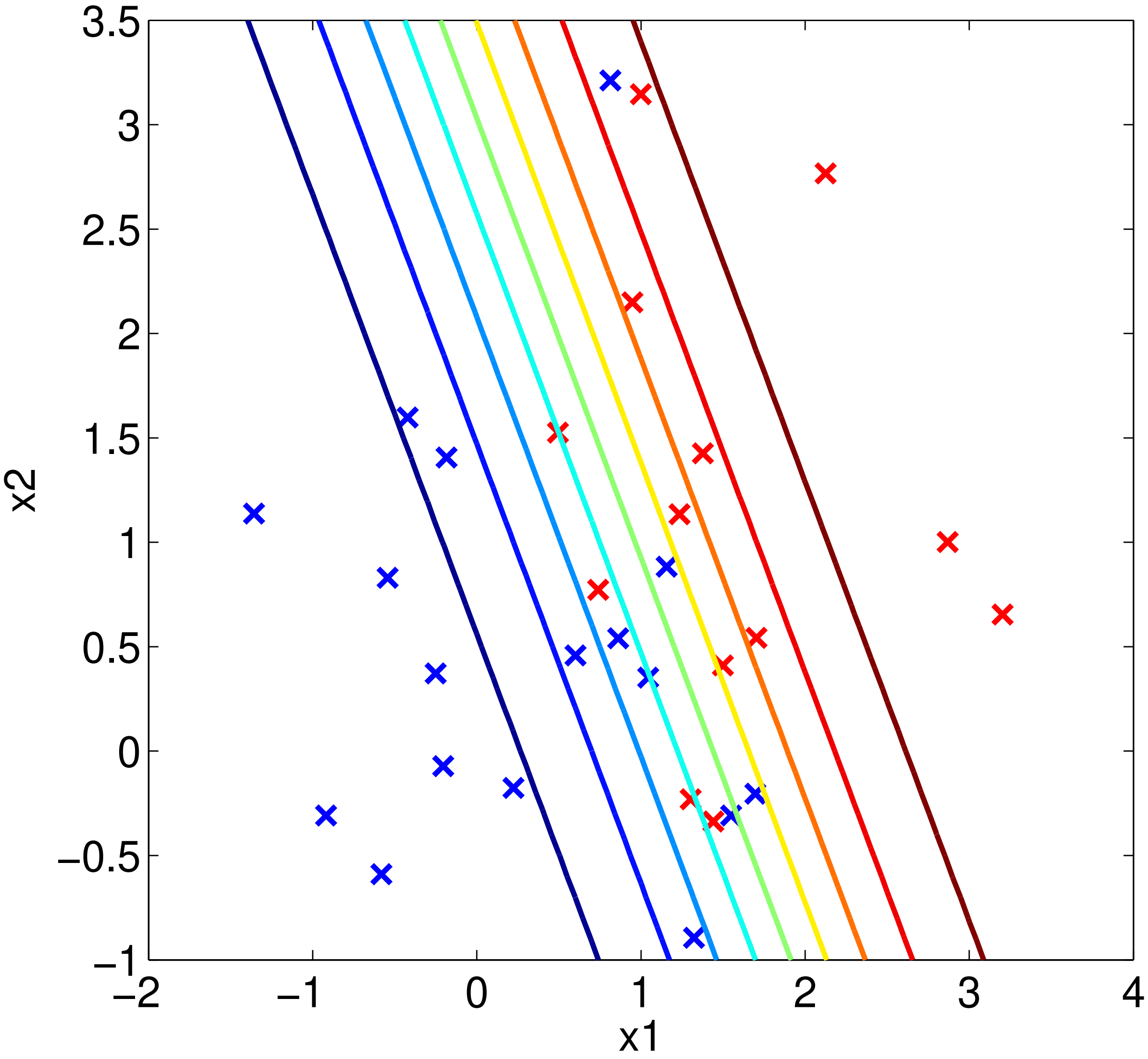
Artifical Example: Data



Artifical Example: Class conditional densities



Artifical Example: Posterior $P(y = 1|\mathbf{x})$



The decision boundary

- The *decision boundary* is the set of $\mathbf{x}$ where $P(y = 1|\mathbf{x}) = P(y = 0|\mathbf{x}) = \frac{1}{2}$.
- What is the decision boundary for LDA?
- More generally, what is the form of the function $P(y = 1|\mathbf{x})$?

LDA decision boundary

The decision boundary is defined by $P(y=0|\mathbf{x}) = P(y=1|\mathbf{x})$. Applying Bayes's rule to boths sides, we get:

$$\begin{aligned}
\frac{P(\mathbf{x}|y=0)P(y=0)}{P(\mathbf{x})} &= \frac{P(\mathbf{x}|y=1)P(y=1)}{P(\mathbf{x})} \\
\Rightarrow P(\mathbf{x}|y=0)P(y=0) &= P(\mathbf{x}|y=1)P(y=1) \\
\Rightarrow \frac{\exp(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0))P(y=0)}{\sqrt{(2\pi)^n|\Sigma|}} &= \frac{\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)}{\sqrt{(2\pi)^n|\Sigma|}}
\end{aligned}$$

From here, it gets rather ugly to put on slides. The denominators on both sides cancel. The $P(y)$ terms can be brought inside the $\exp$ as $+\log P(y)$. Then we take the $\log$ of both sides. The key point is that $x^T\Sigma{-}1x$ terms on each side cancel, and we are left with something of the general form $a^Tx + b = 0$. The decision boundary is thus a linear subspace of $\Re^n$ of dimension $n-1$. That is, a line when $n=2$, a plane when $n=3$, etc. This boundary separate the halves of the input space where $P(y=0|\mathbf{x}) > P(y=1|\mathbf{x})$ from where $P(y=0|\mathbf{x}) < P(y=1|\mathbf{x})$.

LDA posterior function

What is the form of the posterior $P(y = 1|\mathbf{x})$?

$$\begin{aligned}
P(y=1|\mathbf{x}) &= \frac{P(\mathbf{x}|y=1)P(y=1)}{P(\mathbf{x})} \\
&= \frac{P(\mathbf{x}|y=1)P(y=1)}{P(\mathbf{x}|y=0)P(y=0)+P(\mathbf{x}|y=1)P(y=1)} \\
&= \left(1+\frac{P(\mathbf{x}|y=0)P(y=0)}{P(\mathbf{x}|y=1)P(y=1)}\right)^{-1} \\
&= \left(1+\frac{\exp(-\frac{1}{2}(\mathbf{x}-\mu_0)^T\Sigma^{-1}(\mathbf{x}-\mu_0))P(y=0)}{\exp(-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma^{-1}(\mathbf{x}-\mu_1))P(y=1)}\right)^{-1}
\end{aligned}$$

With further algebraic simplification, one can get to

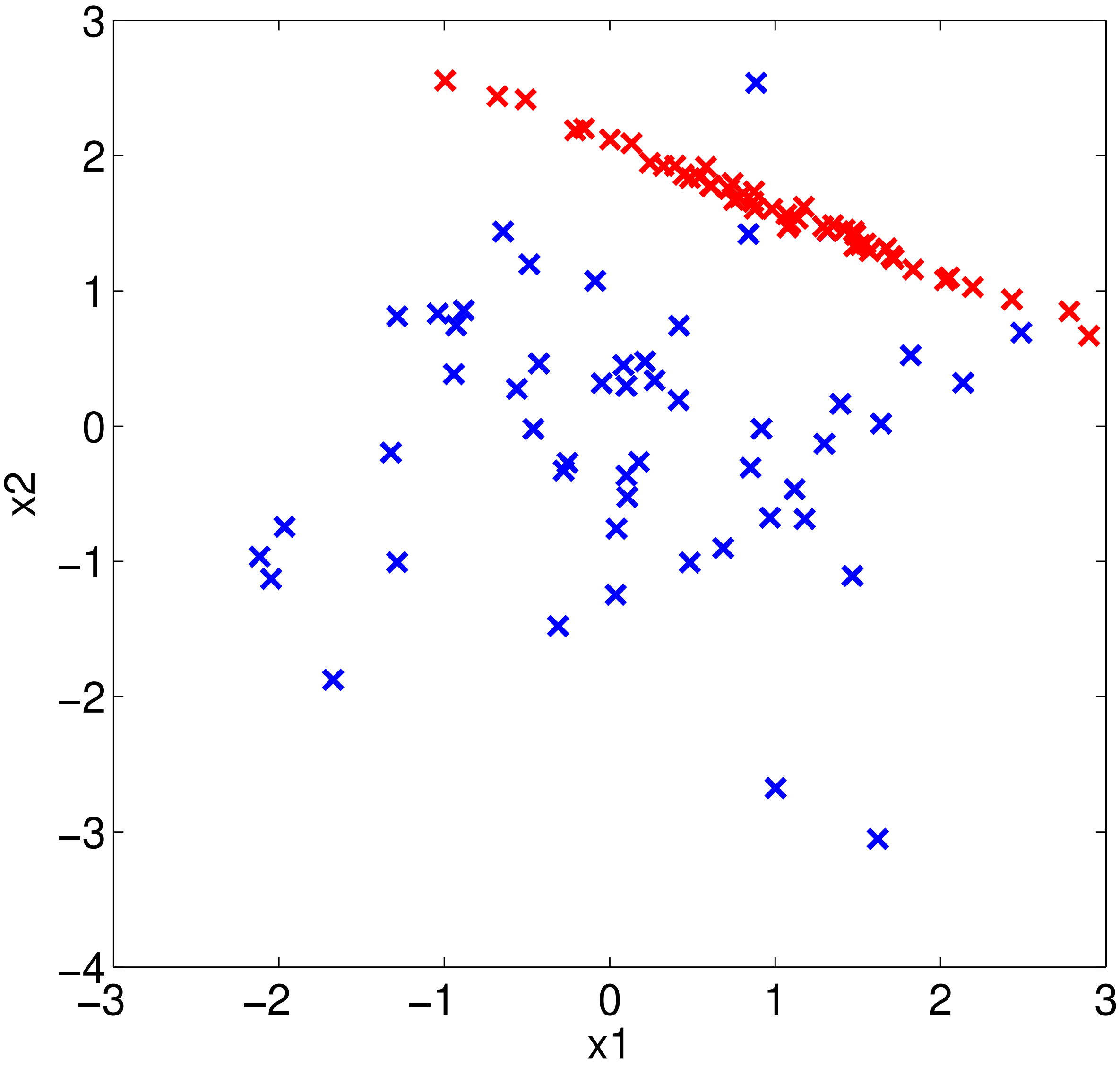
$$P(y=1|\mathbf{x}) = (1+\exp(\mathbf{w}^T\mathbf{x}+w_0))^{-1} = \sigma(\mathbf{w}^T\mathbf{x}+w_0)$$

for appropriate vector $\mathbf{w}$ and constant w_0. It's a logistic classifier!

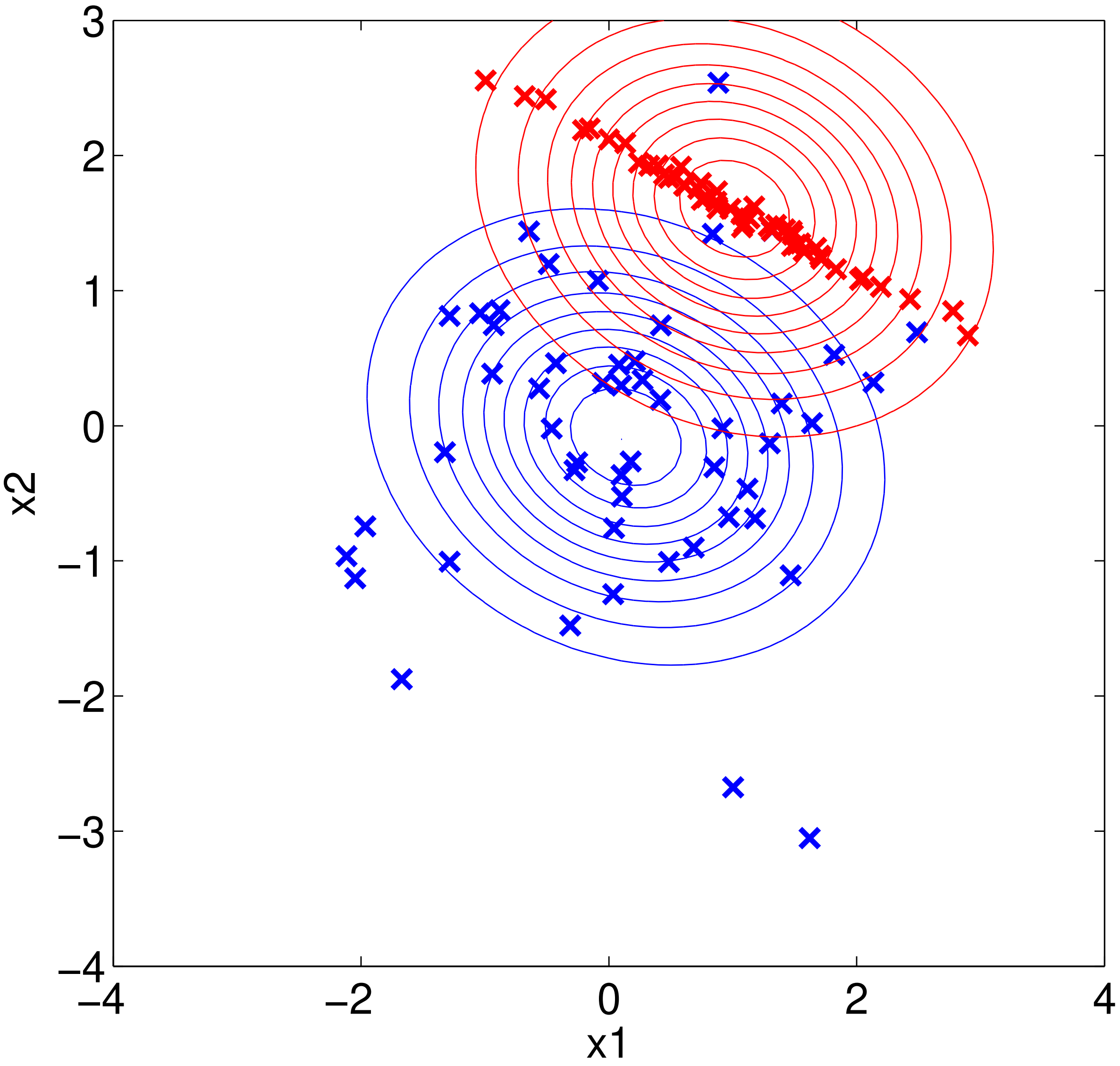
LDA posterior function (II)

- $P(y = 1|\mathbf{x})$ is of the form $\sigma(\mathbf{w}^T\mathbf{x})$, where σ is the logistic function.
- This does not mean, however, that $\mathbf{w}$ are the same weights as would be obtained by logistic regression!
- Particularly when the covariances of the classes are not really the same, LDA may perform poorly – worse than logistic regression would.

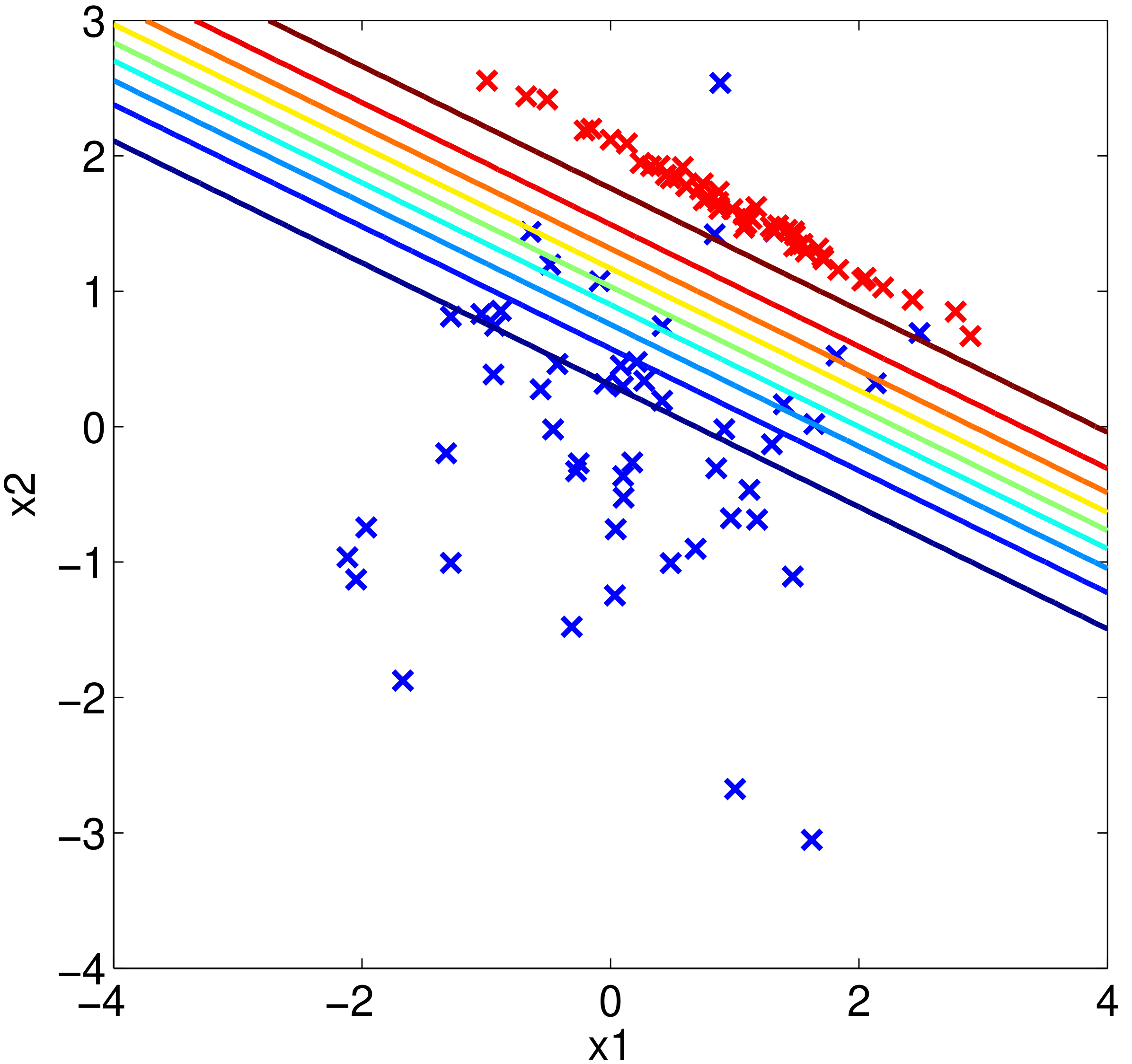
LDA on another example



LDA on another example



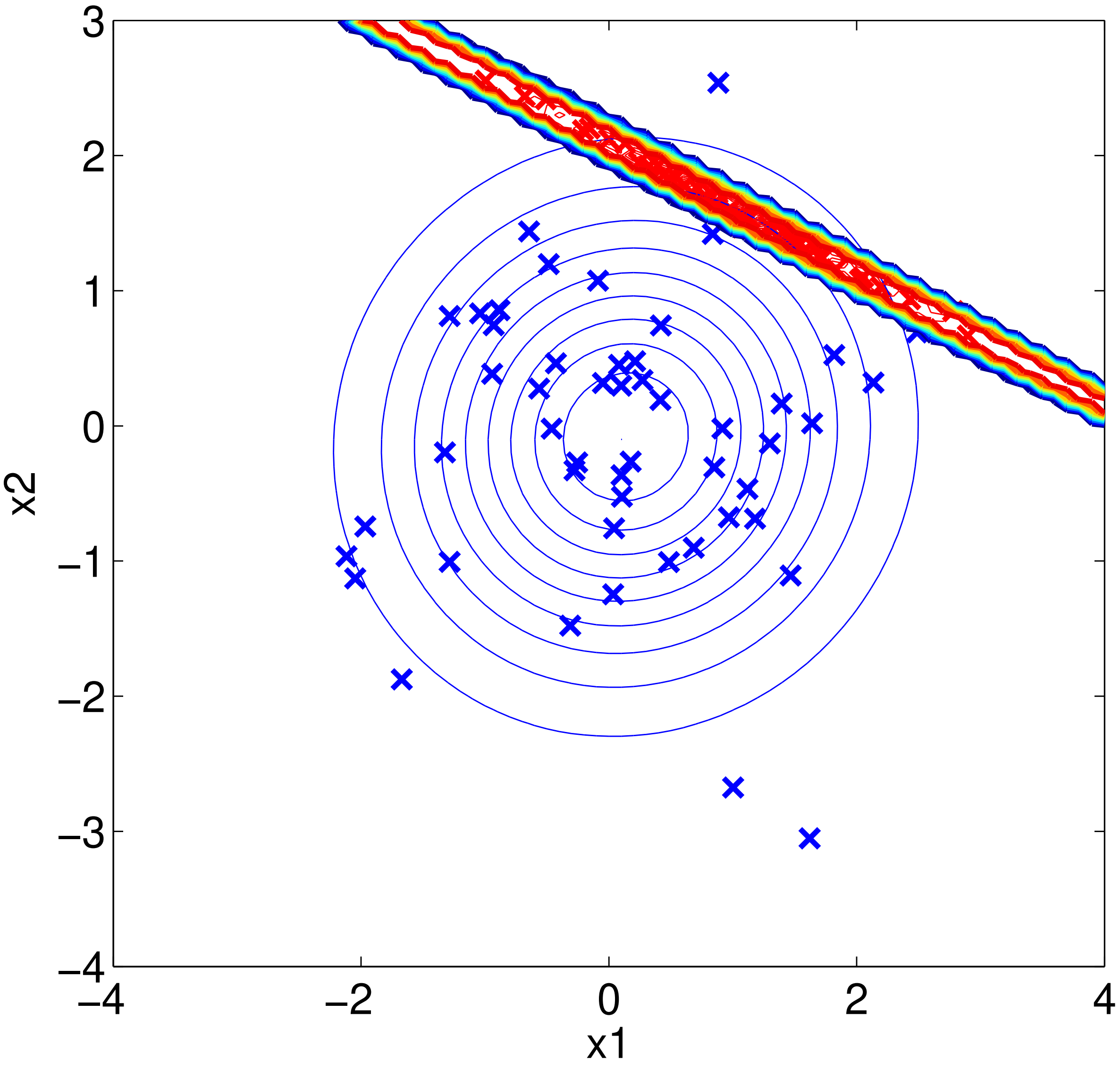
LDA on another example



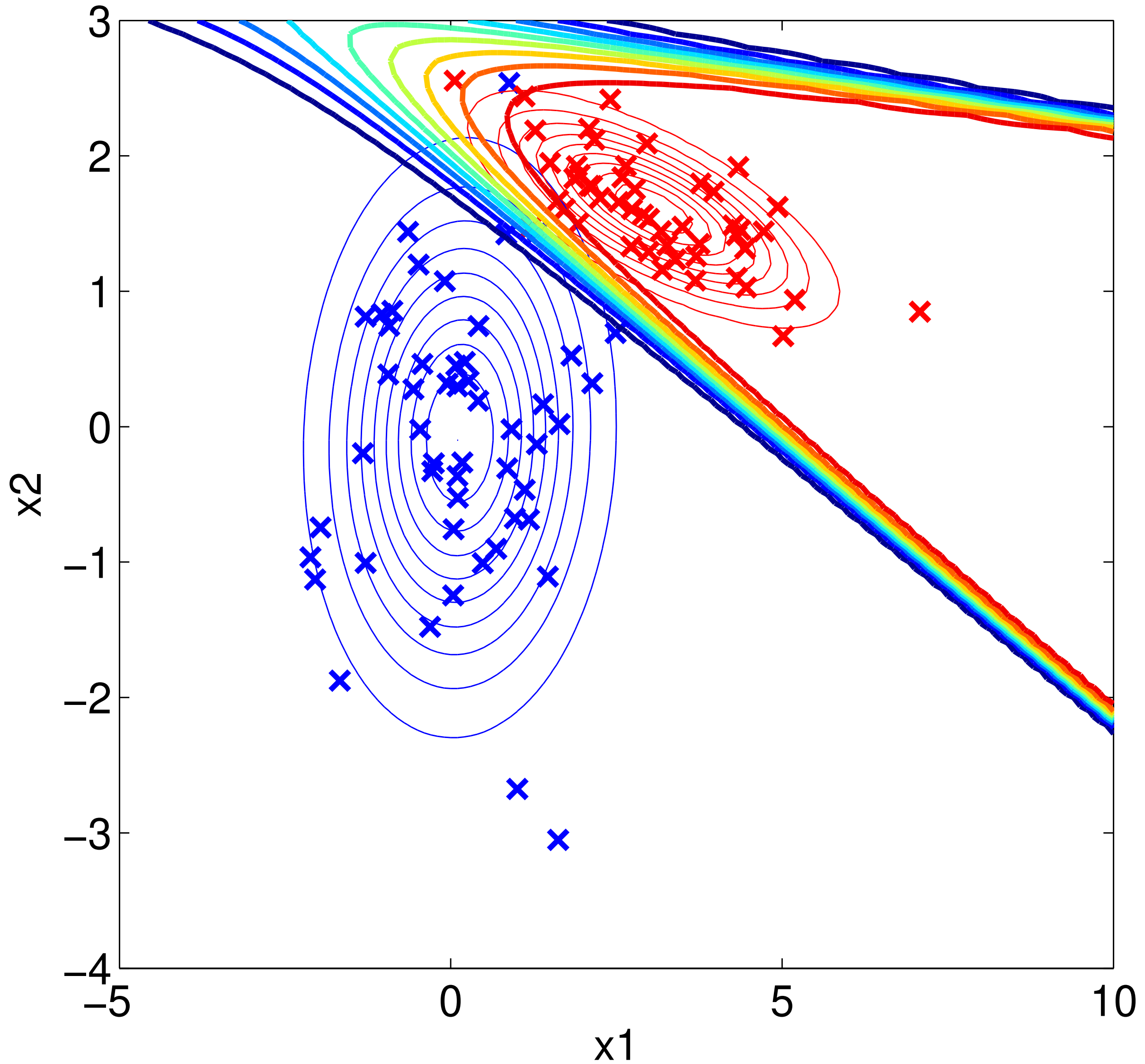
Allowing different covariance matrices

- *Quadratic discriminant analysis* (QDA), is GDA when each class y is allowed its own covariance matrix Σ_y.
  (Guess why it's called *quadratic!*)
- Each μ_y and Σ_y estimated in the standard, maximum likelihood way.
- Can perform better than LDA when classes have much different covariance structures.
- But, more parameters are being fit, so one must be careful of overfitting.

QDA example



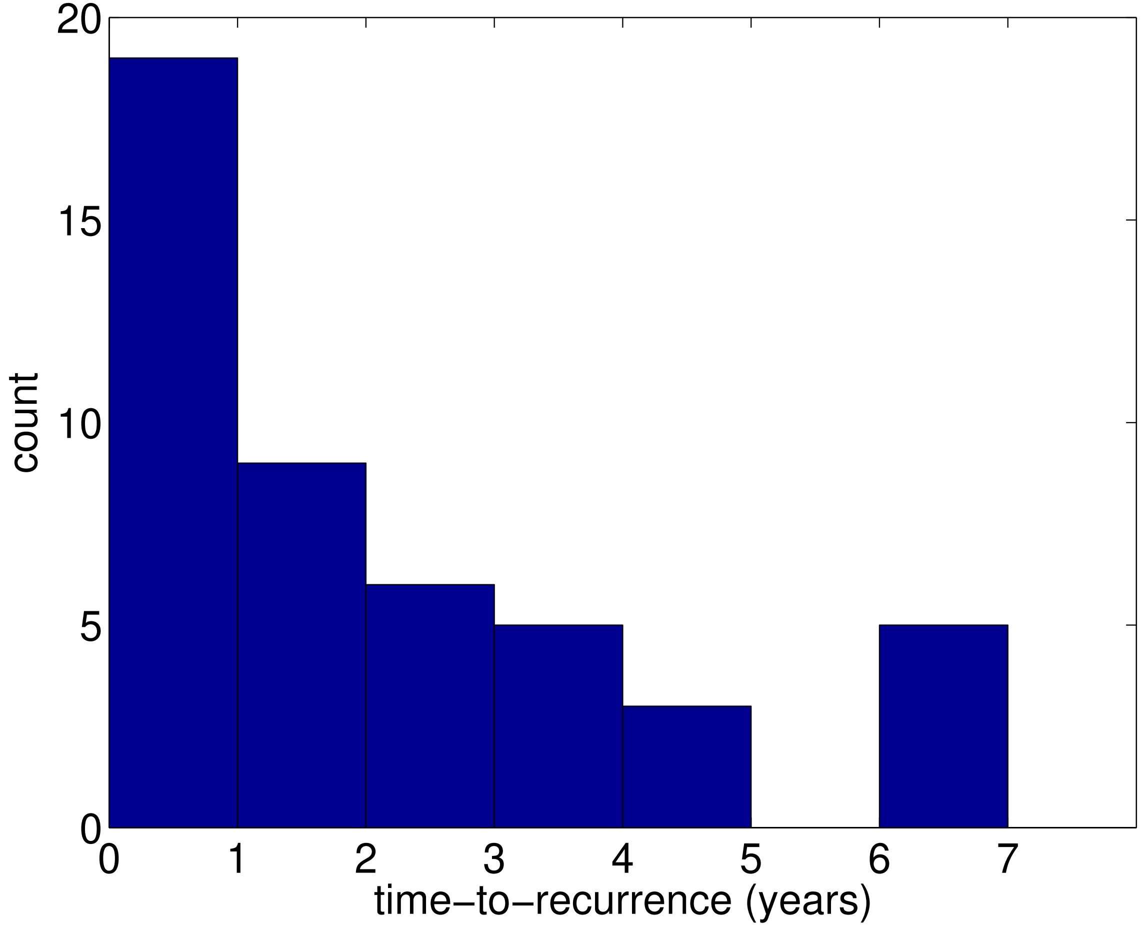
Another QDA example



Not everything is Gaussian!

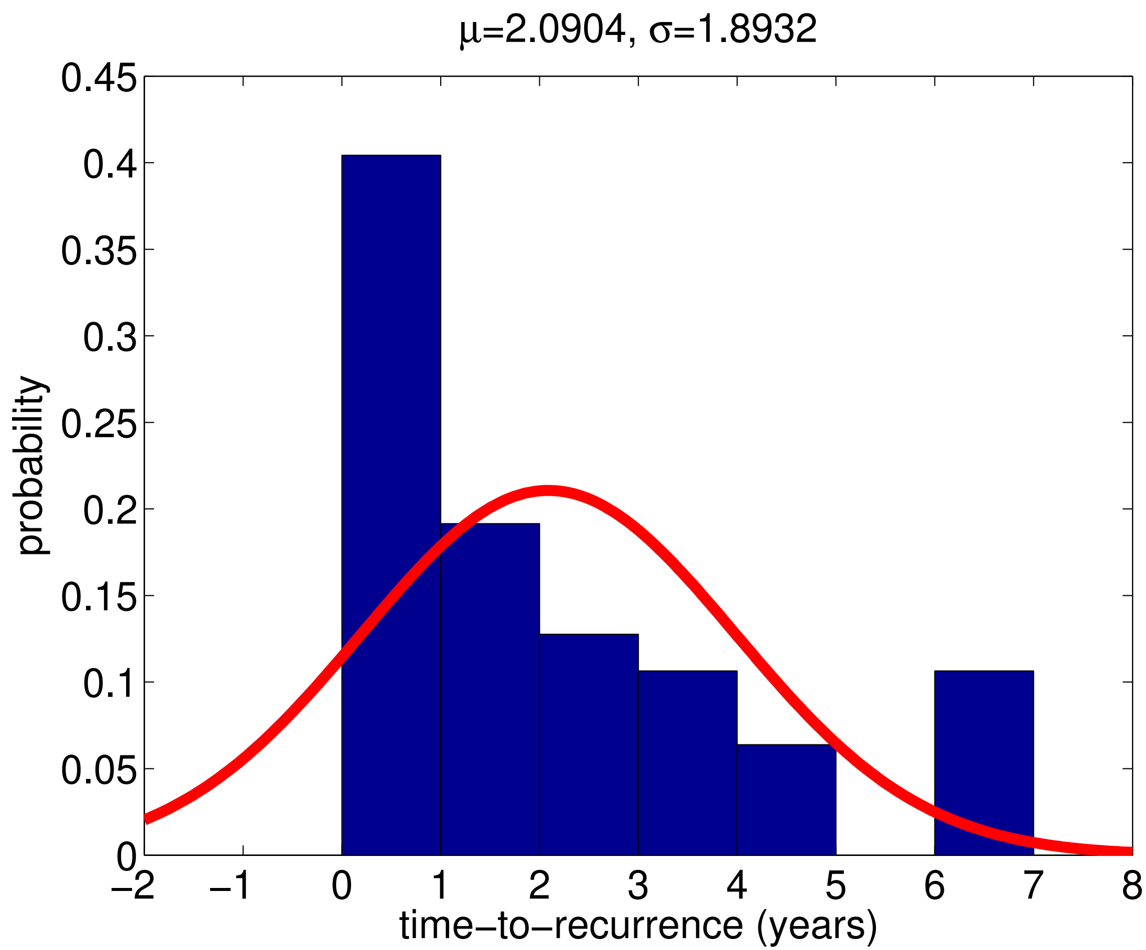
Time-to-recurrence, for those patients who have recurrences:

patient	t-to-r (months)
1	27
2	77
3	77
4	36
5	10
6	10
7	9
⋮	⋮



Not everything is Gaussian! (II)

- Maximum likelihood Gassian fit captures the data poorly.



Other distributions & densities

- We have shown the maximum likelihood estimates for Gaussian distributions
- Estimate for other distributions can be derived similarly. E.g.:
  - Exponential
  - Uniform
  - Lognormal
  - Bernoulli
  - Discrete
  - ...

E.g., maximum likelihood estimate for the exponential

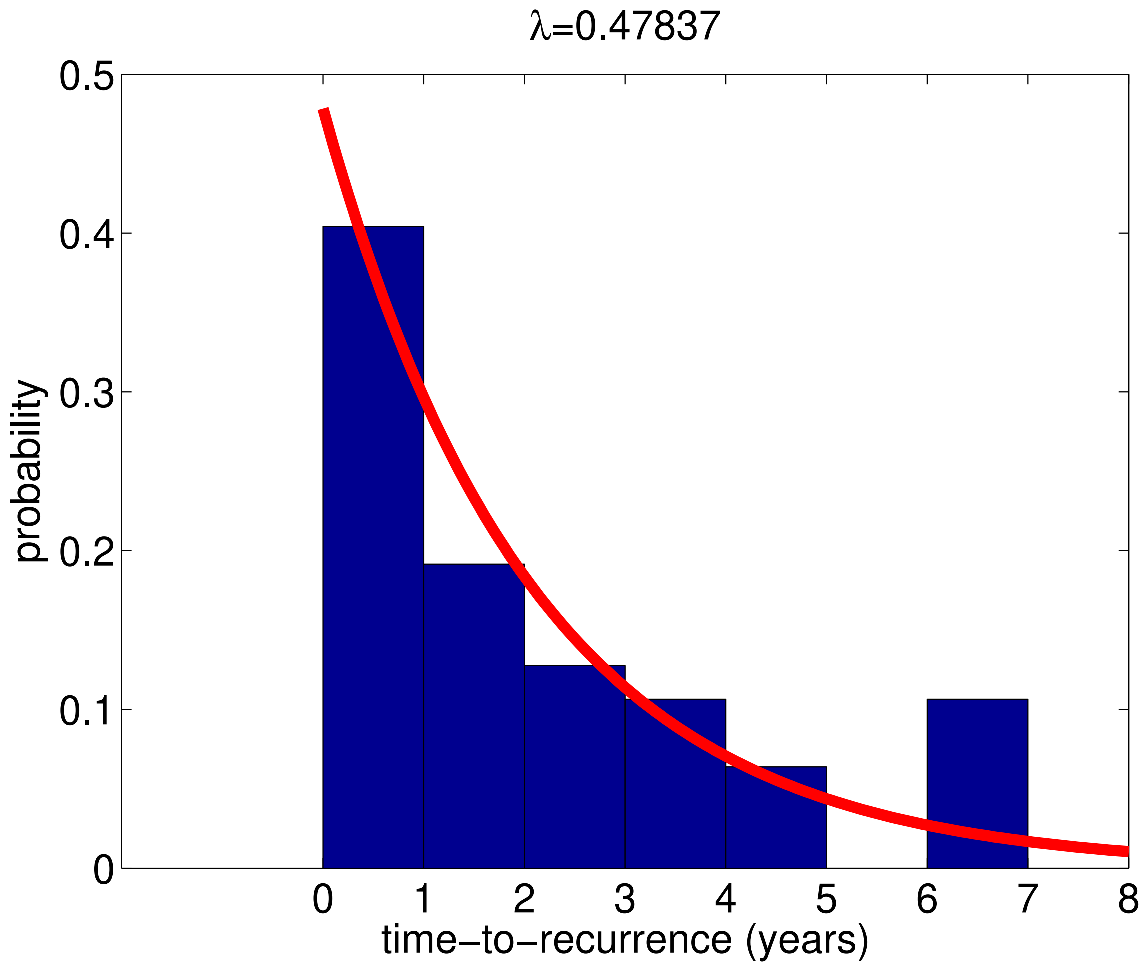
- Suppose nonnegative univariate data $x_1, x_2, \ldots, x_m$.
- Suppose i.i.d. from $P(x) = \lambda e^{-\lambda x}$.
- Maximum likelihood of data w.r.t. λ:

$$\begin{aligned}
& \arg\max_\lambda l(x_1, \ldots, x_m | \lambda) \\
= \quad & \arg\max_\lambda \Pi_{i=1}^m \lambda e^{-\lambda x_i} \\
= \quad & \arg\max_\lambda \log \Pi_{i=1}^m \lambda e^{-\lambda x_i} \\
= \quad & \arg\max_\lambda m \log(\lambda) - \lambda \sum_{i=1}^m x_i
\end{aligned}$$

- Setting the derivative w.r.t. λ equal to zero, we obtain $\lambda = \frac{m}{\sum_i x_i}$, the inverse of the sample average.

Example – Time-to-recurrence data

- $P(x) = \lambda e^{-\lambda x}$
- $\lambda = \frac{m}{\sum_i x_i}$



Estimating discrete distributions

Estimating discrete distributions comes up in several ways:

- Estimating the prior probabilities of different classes. $P(y = 0)$ and $P(y = 1)$, for example.
- Estimating the class-conditional probabilities of a discrete feature. For example: time-to-recurrence (in months).
- Estimating the class-conditional probabilities of a discretized feature. For example:

	recur	not recur
cell size > 17.4	31	16
cell size ≤ 17.4	66	85

Estimating discrete distributions (II)

- We assume data $x_1, x_2, \ldots, x_m \in \mathcal{X}$ where $\mathcal{X}$ is a finite set.
- Without loss of generality, suppose that $\mathcal{X} = \{1, 2, \ldots, k\}$.
- Let m_j be the number of data items that are j.
- The then maximum likelihood estimates are just the empirical frequencies:

$$P(x = j) = \frac{m_j}{m}$$

Proof for binary data

. . . is straightforward. Just write down the likelihood of the data, as a function of $p_1 = P(x = 1)$ and $p_2 = P(x = 2) = 1 - p_1$, differentiate w.r.t. p_1, set equal to zero, and solve.

Proof for the general case

- Let p_i be the probability $P(x = i)$, for $i \in \{1, \ldots, k\}$.
- For data $x_1, x_2, \ldots, x_m$, the likelihood is

$$l(x_1, \ldots, x_m | p_1, \ldots, p_k) = \Pi_{i=1}^{m} p_{x_i}$$

- Collecting by possible values of x, we obtain

$$l(x_1, \ldots, x_m | p_1, \ldots, p_k) = \Pi_{j=1}^{k} p_j^{m_j}$$

  where m_j is the number of times j appears in the data set.
- The log likelihood is thus

$$\log l(x_1, \ldots, x_m | p_1, \ldots, p_k) = \sum_{j=1}^{k} m_j \log p_j$$

- We want to maximize this w.r.t. the p_j, but the p_j are constrained. They must be between zero and one, and they must sum to one...