Lecture 9: Density estimation I

- **Overview**
  - Parametric Vs. Non-parametric methods
- **Maximum Likelihood parameter estimation**
- **Non-parametric density estimation**
  - Histogram
  - K Nearest Neighbor

Overview (1)

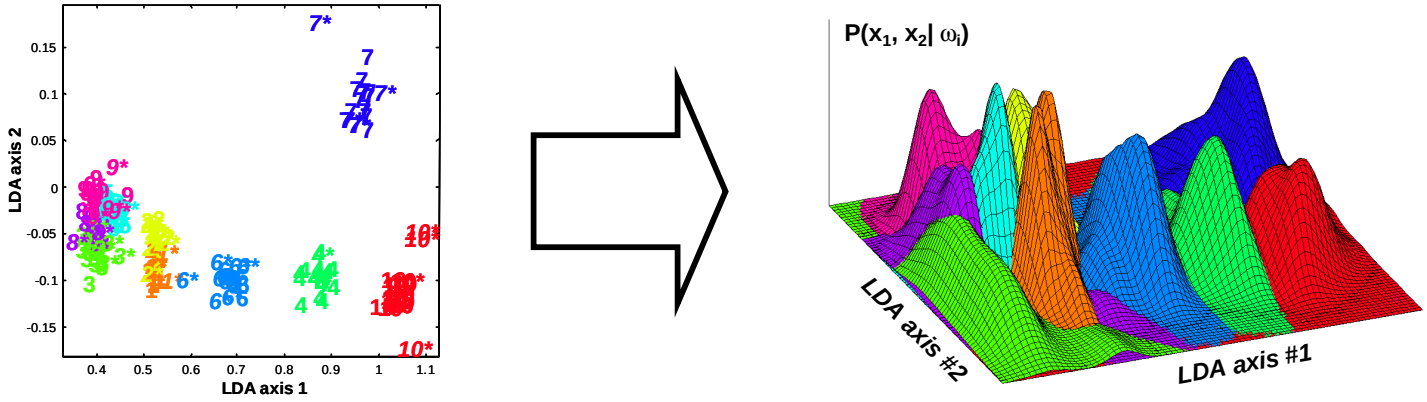
- **At the risk of sounding repetitive, from our decision-theoretic discussion we concluded that the optimal classifier could be expressed as a family of discriminant functions**

$$g_i(x) = P(w_i \mid x) = \frac{P(x \mid w_i)P(w_i)}{P(x)} \propto P(x \mid w_i)P(w_i)$$

  - and the decision rule was

$$\text{choose } \omega_i \text{ if } g_i(x) > g_j(x)\ \forall j \neq i$$

  - In order to build these discriminant functions we need to estimate both prior $P(\omega_i)$ and likelihood $P(x|\omega_i)$
    - The prior will normally be derived from knowledge about the problem, but the estimation of the likelihood is not that easy
  - The objective of the next two lectures is to present a group of techniques to estimate the likelihood density functions $P(x|\omega_i)$ and, ultimately, build the discriminant functions



Overview (2)

- **Density estimation is the problem of modeling a density P(x) given a finite number of data points $x^{(N}$ drawn from that density function**
  - For our purposes we will have a finite number of examples from each class $x^{(Ni}$ (i=1…C) and will model each of the likelihoods $P(x|\omega_i)$ separately
    - From now on we will omit the class label for simplicity, but always keep in mind that we are estimating a class-conditional density
- **There are two basic approaches to perform density estimation**
  - Parametric: a given form for the density function is assumed (i.e., Gaussian) and the parameters of the function (i.e., mean and variance) are then optimized by fitting the model to the data set
    - Parametric density estimation is normally referred to as Parameter Estimation
      - When you compute the sample mean or sample covariance matrix, you are doing parameter estimation in the Maximum Likelihood sense, as we will see in the next few slides
  - Non-parametric: no functional form for the density function is assumed, and the density estimates is driven entirely by the data
    - As an example, when you compute a histogram, you are doing non-parametric density estimation.
    - Other techniques we will cover are K Nearest Neighbor and Kernel (non-parametric) density estimation

Maximum Likelihood parameter estimation

- **Consider a p.d.f. P(x) which depends on a set of parameters $\theta=(\theta_1,\ldots,\theta_M)$**
  - To make the dependency more explicit we will write $P(x|\theta)$
- **Along with this model we have a data set of N vectors $X=\{x^{(1}, x^{(2}, \ldots x^{(N}\}$ from which we want to estimate the parameters $\theta=(\theta_1, \ldots \theta_M)$**
  - **Again**: in a pattern recognition problem these samples X will come from a given class ω_i and we will estimate $P(x|\theta,\omega_i)$ (we omit the dependency on the class for simplicity)
- **If the vectors in the data set are drawn independently from the distribution $P(x|\theta)$, then the joint probability density of the entire data set is given by**

$$P(X|\theta)=\prod_{n=1}^{N}P(x^{(n}|\theta)$$

  - $P(X|\theta)$ is also a likelihood function (the likelihood of parameters θ given the data set X)
- **We will seek the set of parameters θ_{ML}, that maximize this likelihood, and will call it the Maximum Likelihood estimate of θ**
  - Intuitively, θ_{ML} corresponds to the value of θ that agrees best with the observed data
  - Other parameter estimation criteria exist, but they are beyond the scope of our discussion
- **For analytical purposes it is easier to work with the logarithm of the likelihood, so we define the log-likelihood as**

$$\ell(\theta)=\log(P(X|\theta))=\log\left(\prod_{n=1}^{N}P(x^{(n}|\theta)\right)=\sum_{n=1}^{N}\log(P(x^{(n}|\theta))$$

  - And the ML estimate of the parameter is found by finding the zeros of its gradient vector:

$$\nabla_\theta\ell(\theta)=\sum_{n=1}^{N}\nabla_\theta\log(P(x^{(n}|\theta))=0\ \Rightarrow\ \theta_{ML}$$

ML parameter estimation: Gaussian case (1)

- **The most typical situation will involve the estimation of the parameters of a Gaussian distribution**
- **Let's find what these ML estimates become for the univariate case**
  - In this case $\theta_1=\mu$ and $\theta_2=\sigma$ and the log-likelihood becomes

$$\ell(\theta)=\log\left(\prod_{n=1}^{N}\frac{1}{\sqrt{2\pi}\theta_2}e^{-\frac{1}{2\theta_2^2}\left(x^{(n}-\theta_1\right)^2}\right)=\sum_{n=1}^{N}\left[-\frac{1}{2}\log\left(2\pi\theta_2^2\right)-\frac{1}{2\theta_2^2}\left(x^{(n}-\theta_1\right)^2\right]$$

  - and the gradient is

$$\nabla_\theta \log P(X\mid\theta)=\sum_{n=1}^{N}\begin{bmatrix}\frac{1}{\theta_2^2}\left(x^{(n}-\theta_1\right)\\ -\frac{1}{\theta_2}+\frac{\left(x^{(n}-\theta_1\right)^2}{\theta_2^3}\end{bmatrix}$$

  - Setting the gradient to zero yields the expression for the ML estimates

$$\sum_{n=1}^{N}\frac{1}{\theta_2^2}\left(x^{(k}-\theta_1\right)=0 \quad\Rightarrow\quad \theta_{1,ML}=\frac{1}{N}\sum_{n=1}^{N}x^{(k}=\hat{\mu} \qquad (\hat{\mu}\text{ is called the sample mean})$$

$$-\sum_{n=1}^{N}\frac{1}{\theta_2}+\sum_{n=1}^{N}\frac{\left(x^{(k}-\theta_1\right)^2}{\theta_2^3}=0 \quad\Rightarrow\quad \theta_{2,ML}=\frac{1}{N}\sum_{n=1}^{N}\left(x^{(k}-\hat{\mu}\right)^2=\hat{\sigma} \qquad (\hat{\sigma}\text{ is called the sample variance})$$

  - So we obtain the satisfying result that **the Maximum Likelihood estimates of the mean and the variance are the sample mean and sample variance, respectively**

ML parameter estimation: Gaussian case (2)

- **How good are these estimates? One way to determine it is to find the expected value of the estimate and compare it with the true value**
  - If the expected value of the estimate does not coincide with the true value, the estimate is said to be **biased**
  - We compute the expected value of the sample mean

$$E\left[\hat{\mu}\right] = E\left[\frac{1}{N}\sum_{n=1}^{N} x^{(n}\right] = \frac{1}{N}\sum_{n=1}^{N} E\left[x^{(n}\right] = \frac{1}{N}\sum_{n=1}^{N} \mu = \mu$$

    - So the sample mean is an unbiased estimator of the true mean
  - Computation of the expected value of the sample variance is more elaborate
    - It can be shown that it becomes

$$E\left[\hat{\sigma}^2\right] = \frac{N-1}{N}\sigma^2$$

    - So the sample variance is a biased estimator of the true variance. This is because the sample variance uses the sample mean instead of the true mean in its computation
      - This surprising result does not have many practical implications since, for N large enough, this bias is insignificant
      - On the other hand, if this bias becomes significant, it is only because N is very small and we should not be doing statistics with so few examples in the first place!
  - Similarly, it can be shown that the Maximum Likelihood parameter estimates for the multivariate Gaussian are also the sample mean vector and sample covariance matrix

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x^{(n} = \hat{\mu} \quad \text{and} \quad \Sigma_{ML} = \frac{1}{N}\sum_{n=1}^{N} \left(x^{(n} - \hat{\mu}\right)\left(x^{(n} - \hat{\mu}\right)^T = \hat{\Sigma}$$

    - It can also be shown that the sample mean vector is a unbiased estimator, and that the sample covariance matrix is a biased estimator

$$E\left[\hat{\mu}\right] = \mu, \quad \text{and} \quad E\left[\hat{\Sigma}\right] = \frac{N\text{-}1}{N}\Sigma$$

Non-parametric density estimation, the histogram

- **The basic problem of non-parametric density estimation is straightforward: given a set of examples, model the density function of the data without making any assumptions about the form of the distribution**
- **The simplest form of non-parametric D.E. is the familiar histogram**
  - Divide the sample space into a number of bins and approximate the density at the center of each bin by the fraction of points in the training data that fall into the corresponding bin

$$P_H(x) = \frac{1}{N}\frac{[\text{number of } x^{(n} \text{ in same bin as } x]}{[\text{width of bin containing } x]}$$

    - The histogram requires two "parameters" to be defined: bin width and starting position of the first bin
- **The histogram is a very simple form of D.E., but it has various drawbacks**
  - The final shape of the density estimate depends on the starting point of the bins
    - For multivariate data, the final shape of the density is also affected by the orientation of the bins
  - The discontinuities of the estimate are not due to the underlying density, they are only an artifact of the chosen bin locations
    - These discontinuities make it very difficult, without experience, to grasp the structure of the data
  - A much more serious problem is the curse of dimensionality, since the number of bins grows exponentially with the number of dimensions
    - In high dimensions we would require a very large number of examples or else most of the bins would be empty
- **All these drawbacks make the histogram unsuitable for most practical applications except for rapid visualization of results in one or two dimensions**
  - We do not spend more time looking at the histogram

Non-parametric density estimation, general formulation (1)

- **Before we proceed any further let us return to the basic definition of probability to get a solid idea of what we are trying to accomplish**
  - The probability that a vector x, drawn from a distribution P(x), will fall in a region $\Re$ of the sample space is

$$P = \int_{\Re} P(x')dx'$$

  - Suppose now that N vectors $\{x^{(1}, x^{(2}, \ldots, x^{(N}\}$ are drawn from the distribution. The probability that k of these N vectors fall in $\Re$ is given by the binomial distribution

$$P(k) = \binom{N}{k} P^k (1-P)^{N-k}$$

  - It can be shown (from the properties of the binomial p.m.f.) that the mean and variance of the ratio k/N are

$$E\left[\frac{k}{N}\right] = P \quad \text{and} \quad Var\left[\frac{k}{N}\right] = E\left[\left(\frac{k}{N} - P\right)^2\right] = \frac{P(1-P)}{N}$$

  - Therefore, as $N \to \infty$, the distribution becomes sharper (the variance gets smaller) so we can expect that a good estimate of the probability P can be obtained from the mean fraction of the points that fall within $\Re$

$$P \cong \frac{k}{N}$$

Non-parametric density estimation, general formulation (2)

- On the other hand, if we assume that $\Re$ is so small that P(x) does not vary appreciably within it, then

$$\int_{\Re} P(x')dx' \cong P(x)V$$

  - where V is the volume enclosed by region $\Re$
- Merging with the previous result we obtain

$$\left.\begin{array}{lll} P = \int_{\Re} P(x')dx' & \cong & P(x)V \\ P(x) & \cong & \frac{k}{N} \end{array}\right\} \Rightarrow P(x) \cong \frac{k}{NV}$$

- This estimate becomes more accurate as we increase the number of sample points N and shrink the volume V

- **In practice the value of N is fixed (the total number of examples)**
  - In order to improve the accuracy of the estimate P(x) we could let V to approach zero but then the region $\Re$ would become so small that it would enclose no examples
  - This means that in practice we will have to find a compromise value of the volume V
    - Large enough to include enough examples within $\Re$
    - Small enough to support the assumption that P(x) is constant within $\Re$

Non-parametric density estimation, general formulation (3)

- **So the general expression for non-parametric density estimation is**

$$P(x) \cong \frac{k}{NV} \quad \text{where} \quad \begin{cases} V \text{ is the volume surrounding } x \\ N \text{ is the total number of examples} \\ k \text{ is the number of examples inside } V \end{cases}$$

- **In applying this result to practical density estimation problems there are two basic approaches we can adopt**
  - We can choose a fixed value of k and determine the corresponding volume V from the data. This gives rise to the **k Nearest Neighbor (kNN)** approach
  - We can choose a fixed value of the volume V and determine k from the data. This leads to the methods commonly referred to as **Kernel Density Estimation (KDE)**
- **It can be shown that both kNN and KDE converge to the true probability density as $N \to \infty$, provided that V shrinks with N, and k grows with N appropriately**

kNN Density Estimation

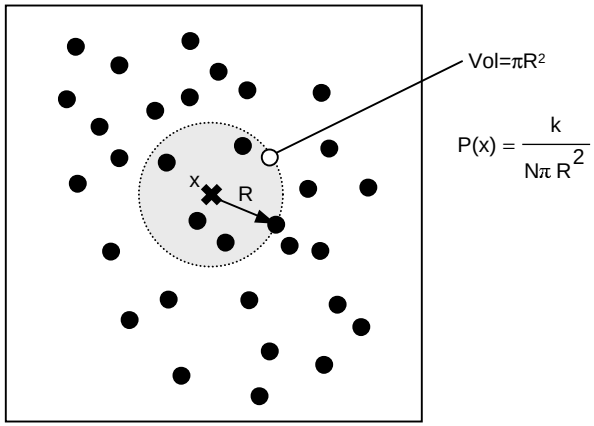
- **In the kNN method we grow the volume surrounding the estimation point x so that it encloses a total of k points**
- **The density estimate then becomes**

$$P(x) \cong \frac{k}{NV} = \frac{k}{N \cdot c_D \cdot R_k^D(x)}$$

- $R_k(x)$ is the distance between the estimation point and its k-th closest neighbor
- c_D is the volume of the unit sphere in D dimensions, which is equal to

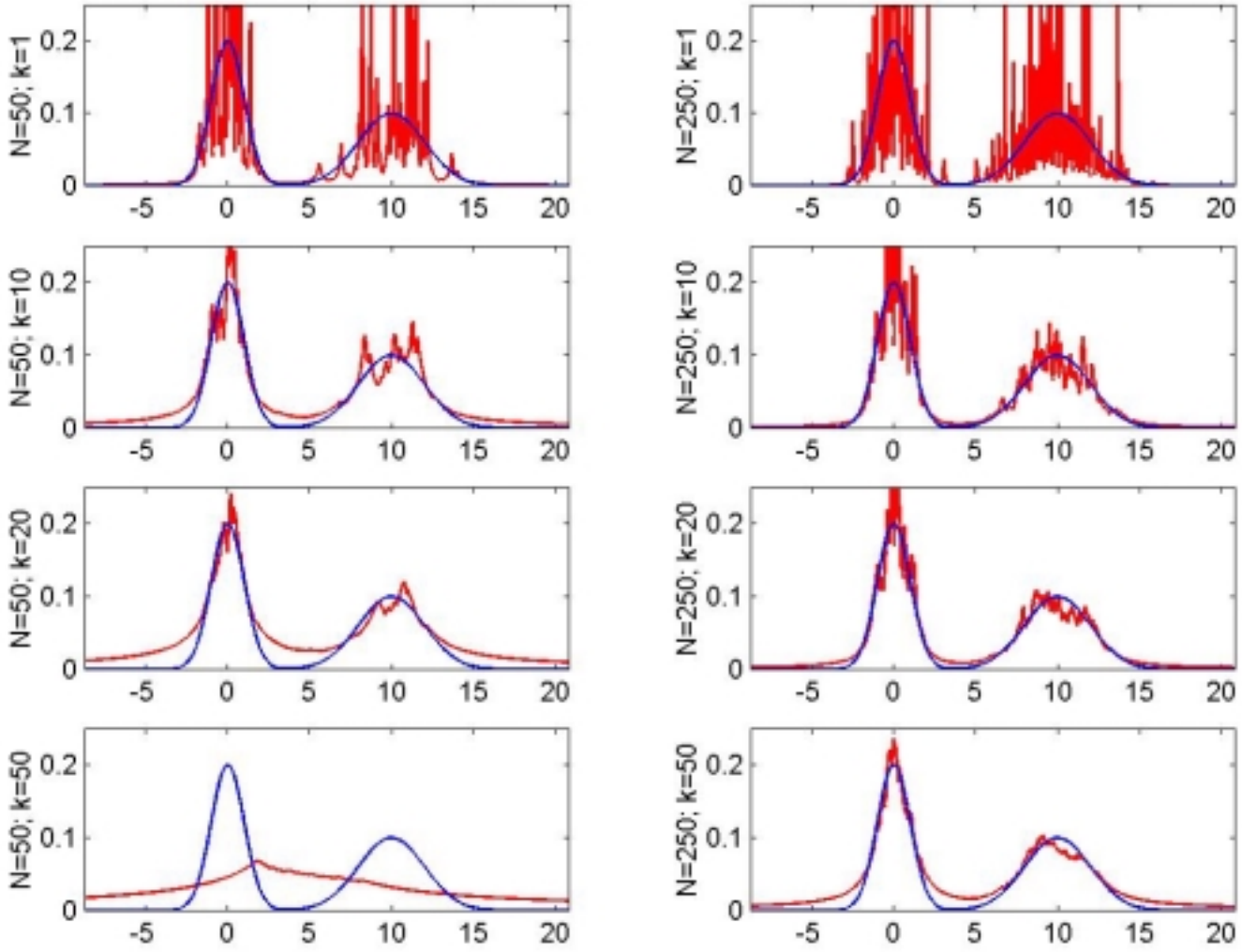
$$c_D = \frac{\pi^{D/2}}{(D/2)!} = \frac{\pi^{D/2}}{\Gamma(D/2+1)}$$

  - Thus $c_1=2$, $c_2=\pi$, $c_3=4\pi/3$ and so on



kNN Density Estimation, example 1

- To illustrate the behavior of kNN we generated several density estimates for a univariate mixture of two Gaussians: P(x)=½N(0,1)+½N(10,4) and several values of N and k



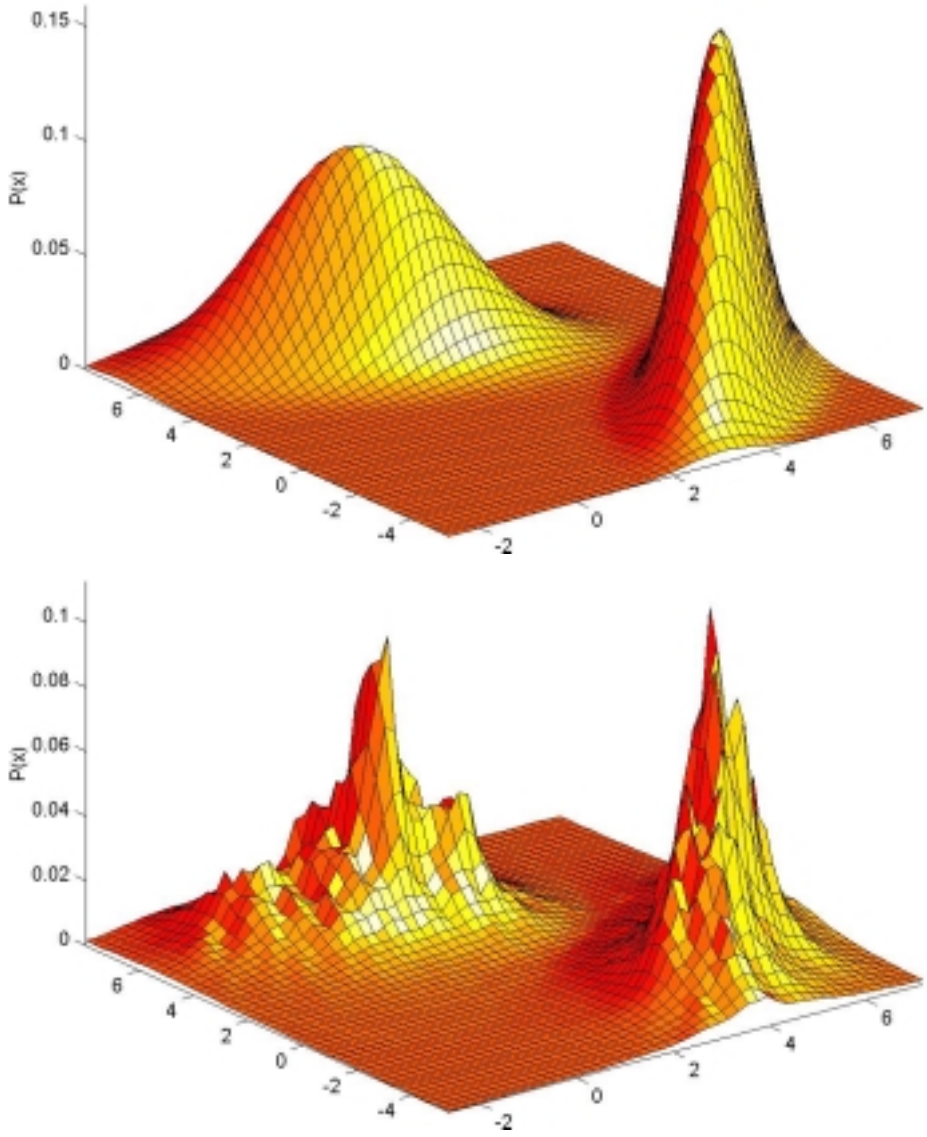
kNN Density Estimation, example 2 (a)

- **The performance of the kNN density estimation technique on two dimensions is illustrated in these figures**
  - The top figure shows the true density, a mixture of two bivariate Gaussians

$$P(x)=\frac{1}{2}N(\mu_1,\Sigma_1)+\frac{1}{2}N(\mu_2,\Sigma_2)$$

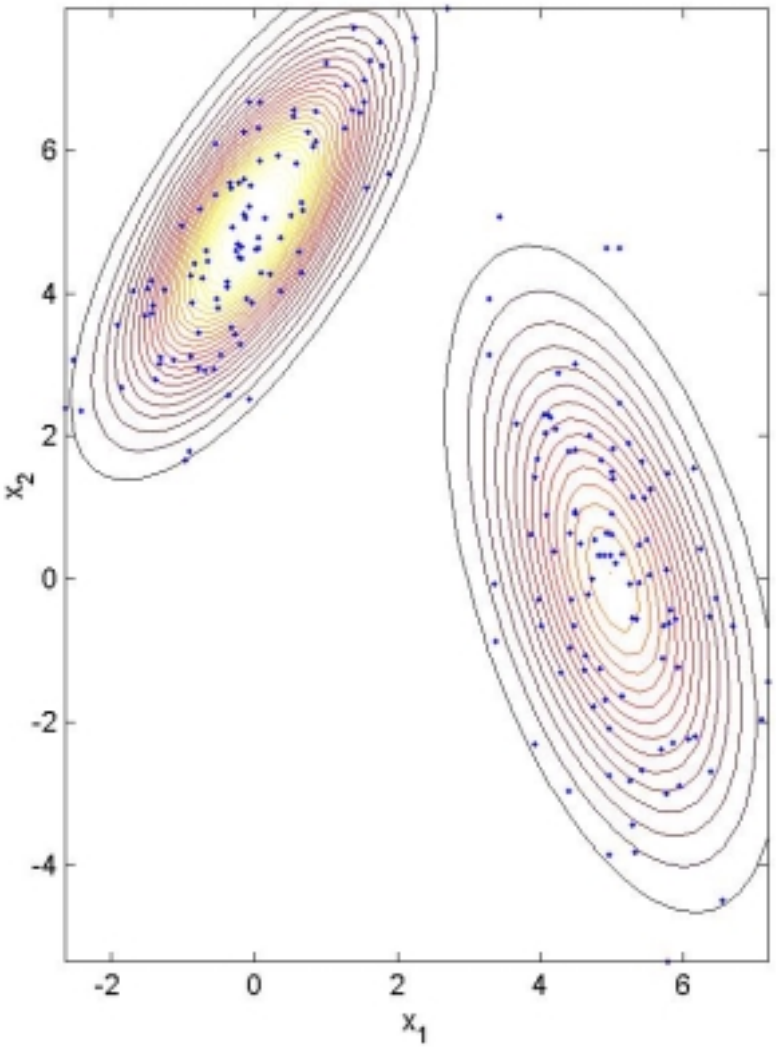
$$\text{with}\begin{cases}\mu_1=[0\ \ 5]^T & \Sigma_1=\begin{bmatrix}1 & 1\\ 1 & 2\end{bmatrix}\\ \mu_2=[5\ \ 0]^T & \Sigma_2=\begin{bmatrix}1 & -1\\ -1 & 4\end{bmatrix}\end{cases}$$

  - The bottom figure shows the density estimate for k=10 neighbors and N=200 examples
  - In the next slide we show the contours of the two distributions overlapped with the training data used to generate the estimate

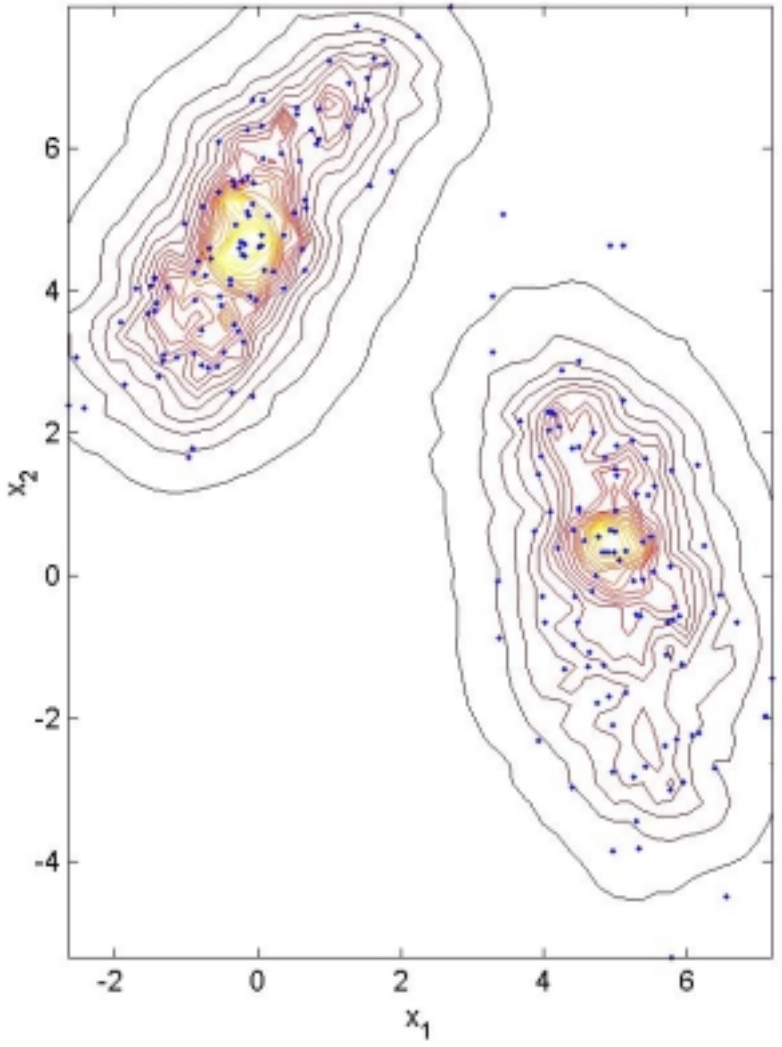


kNN Density Estimation, example 2 (b)

True density contours



kNN density estimate contours



kNN Density Estimation, conclusions

- **The kNN method can be readily used to compute the Bayes classifier**
  - The likelihood functions are estimated by $P(x|\omega_i) = \frac{k_i}{N_i V}$
  - The unconditional density is estimated as $P(x) = \frac{k}{NV}$
  - And similarly the priors can be approximated by $P(\omega_i) = \frac{N_i}{N}$
  - The Bayes classifier becomes $P(w_i|x) = \frac{P(x|w_i)P(w_i)}{P(x)} = \frac{\frac{k_i}{N_i V}\cdot\frac{N_i}{N}}{\frac{k}{NV}} = \frac{k_i}{k}$
    - Notice this is the same decision rule as the k-NNR classifier we derived in the previous lecture!
- **However, the overall estimates that can be obtained with the kNN method are not very satisfactory**
  - The estimates are prone to local noise
  - The method produces estimates with very heavy tails
  - Since the function $R_k(x)$ is not differentiable, the density estimate will have discontinuities
  - The resulting density is not a true probability density since its integral over all the sample space diverges