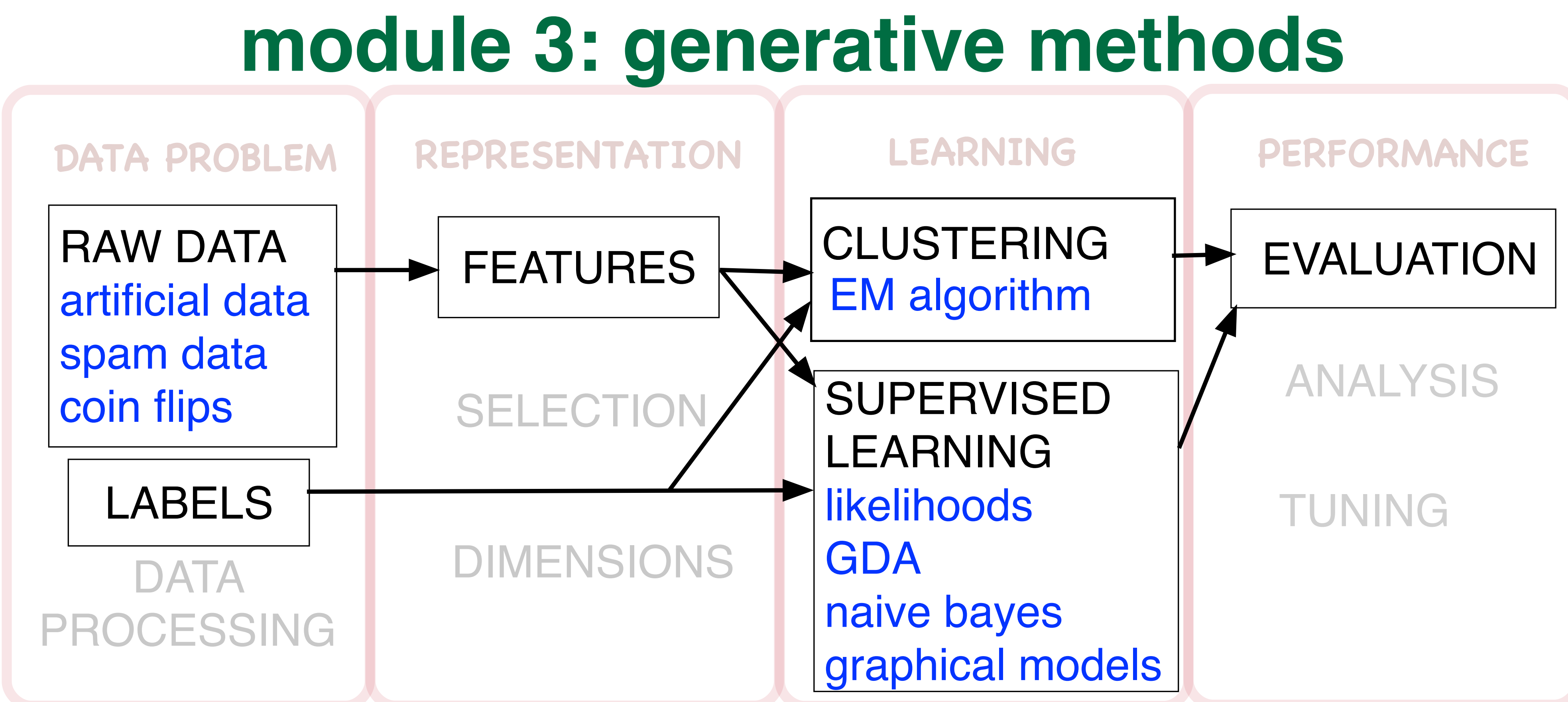
Gaussian Discriminant Analysis

material thanks to Andrew Ng @Stanford

Course Map / module3



- Gaussian Discriminant Analysis

Density Estimation Problem

- $P(y|x) = P(y|x^1,x^2,\ldots,x^d)$ joint (d+1)-dim distribution
- ... actually we cannot estimate this joint
- if each feature has 10 buckets, and we have 100 features (very reasonable assumptions)
- then the joint distribution has 10^{100} cells - impossible

how to get around estimating the joint $P(x^1, x^2, \ldots, x^d | y)$?

- SOLUTION: model/restrict the joint, instead of estimating any possible such joint distribution
  - fore example with a well known parametrized form
  - such as multi-dim gaussian distribution
  - estimate the parameters of the imposed model
- called **Gaussian Discriminant Analysis** (when the model imposed is gaussian)
  - easy to implement due to math tools facilitating gaussian parameters estimation (mean, covariance)
  - multidim implies "covariance" matrix instead of simple variance
  - doesnt fit data in many cases

Gaussian Fit

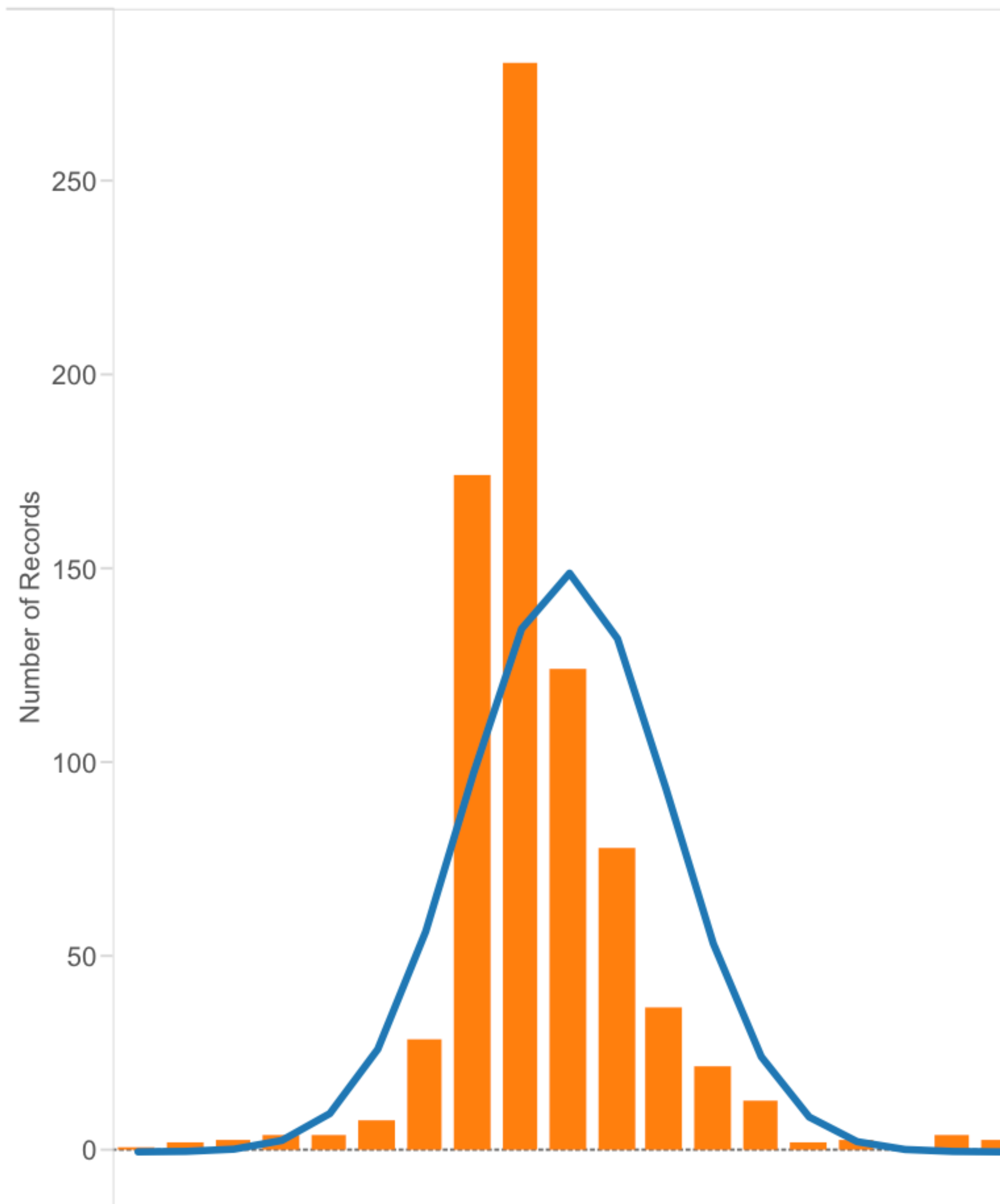
- Idea: fit a parametrized distribution to histogram (density or counts)

- The gaussian (normal) density is controlled by mean and variance

$$P(x|\mu,\sigma^2) = normal(x,\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- the best fit is the one that maximizes likelihood of the data

$$\log L = log\prod_{i=1}^{m} P(x|\mu,\sigma^2) = \sum_{i=1}^{m} logP(x|\mu,\sigma^2)$$



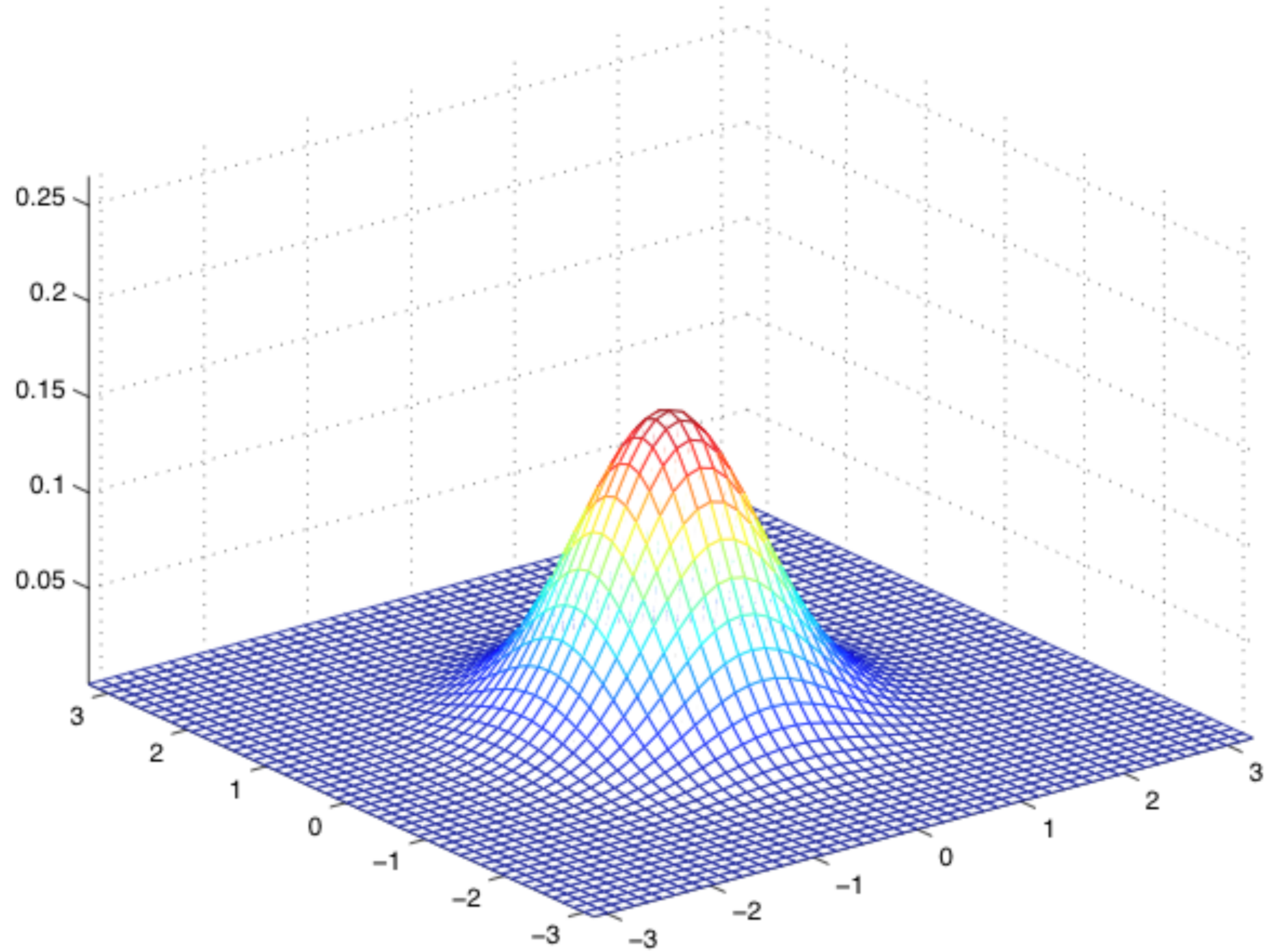
Lets impose a nice probabilistic model

- Multi-variate normal distribution $\theta = (\mu, \Sigma)$

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

  - plotted Σ=identity (or independent variables)

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

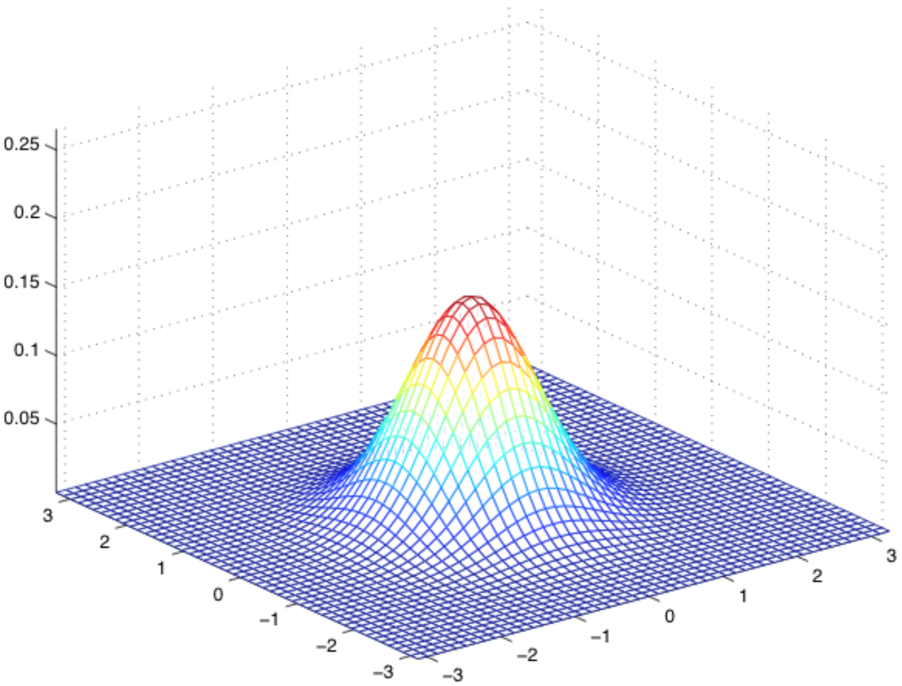


Lets impose a nice probabilistic model

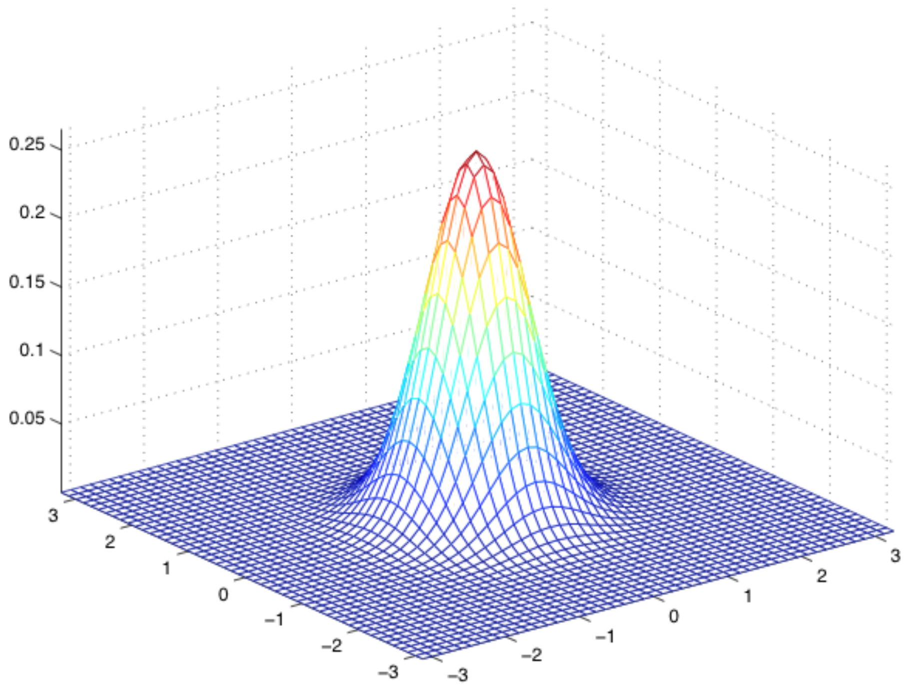
- Multi-variate normal distribution

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

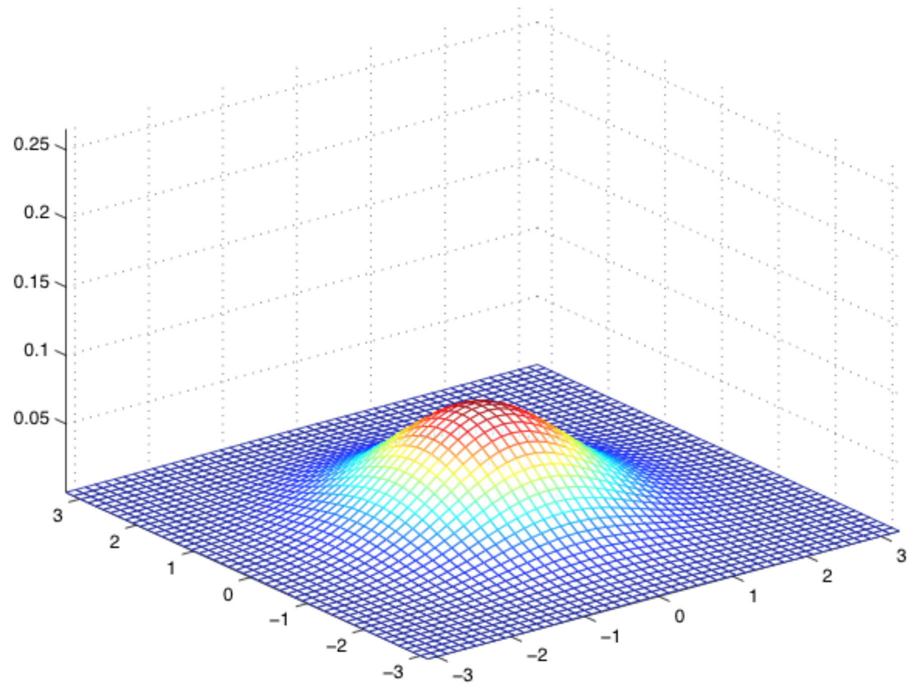
  - plotted Σ=variance only or independent variables



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



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



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

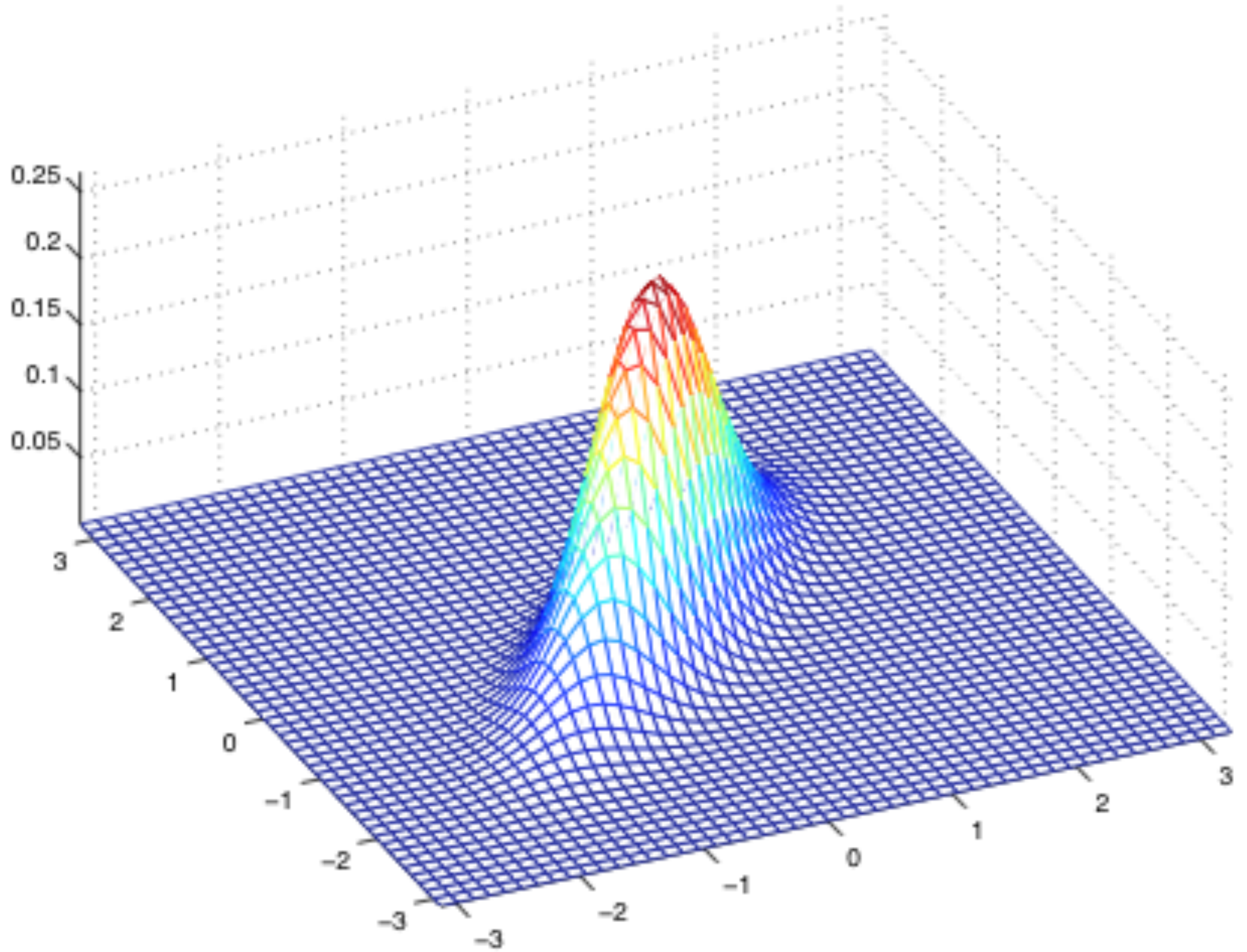
Lets impose a nice probabilistic model

- Multi-variate normal distribution

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

  - plotted $\Sigma \neq$ identity
  - dependent variables

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

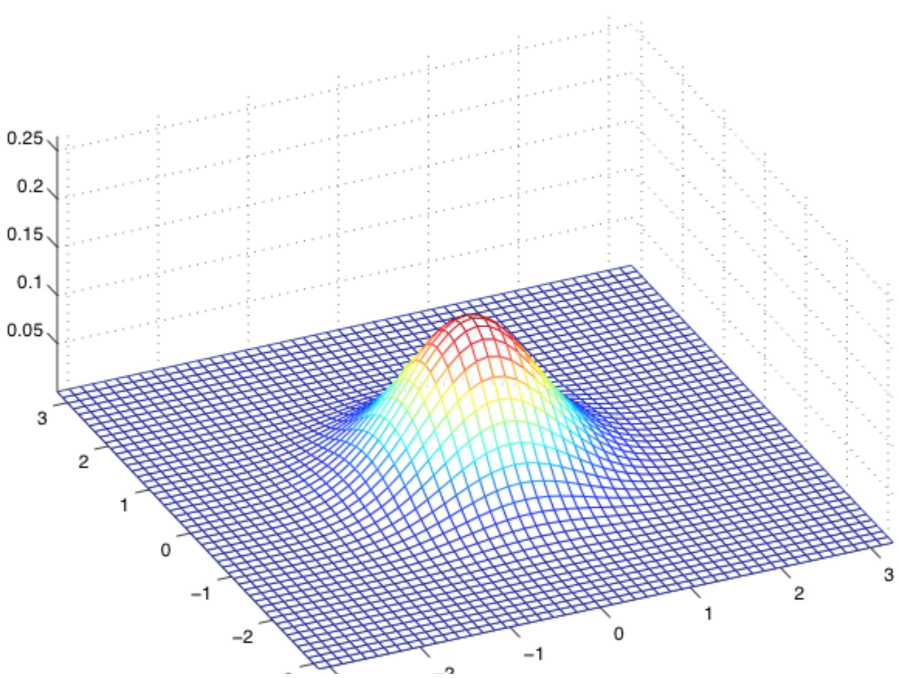


Lets impose a nice probabilistic model

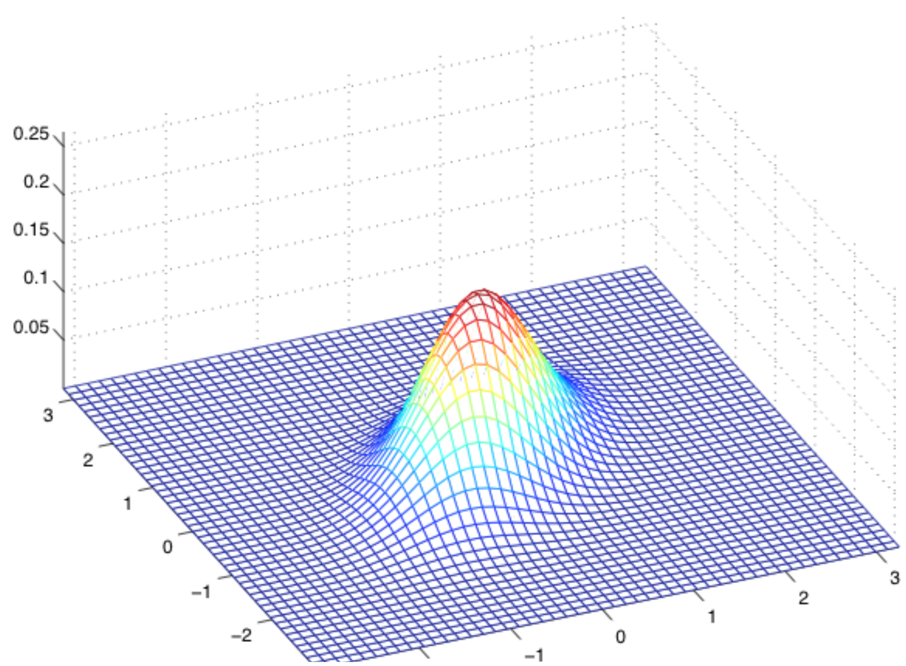
- Multi-variate normal distribution

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

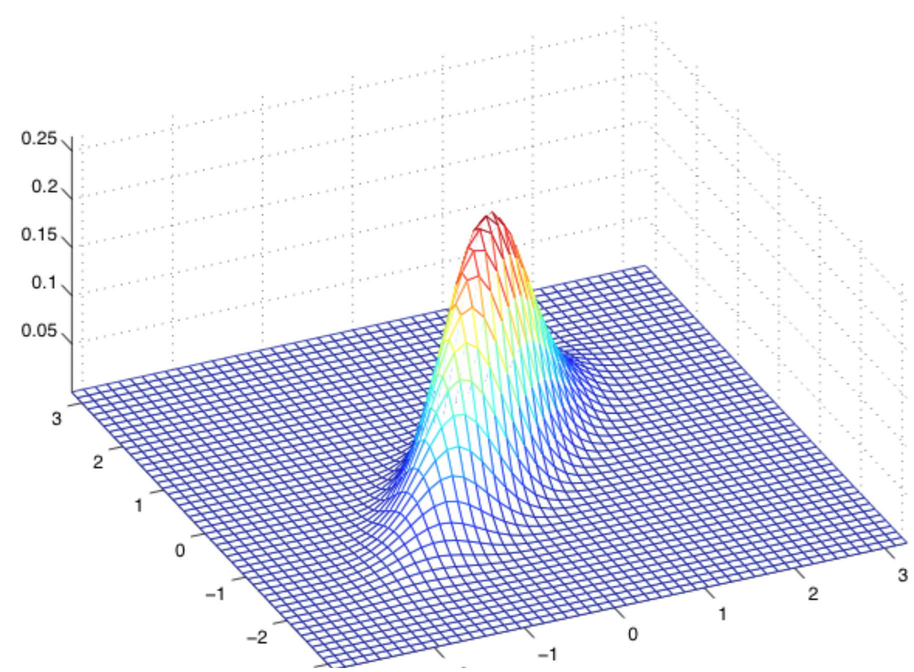
  - Σ≠identity=>dependent variables



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



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



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

GDA Setup

- multi normal density estimation for each y (common Σ)

$$p(x|y=0) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)$$

$$p(x|y=1) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)$$

- log likelihood

$$\ell(\phi,\mu_0,\mu_1,\Sigma) = \log\prod_{i=1}^{m} p(x^{(i)},y^{(i)};\phi,\mu_0,\mu_1,\Sigma)$$

$$= \log\prod_{i=1}^{m} p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma)p(y^{(i)};\phi)$$

GDA parameter solution

- max likelihood for GDA has close form solution!
- can be derived using differentials
  - estimate mean for each class
  - estimate covariance for entire training set
    - or separately for each class
  - no need for Gradient Descent or other optimizers

$$\phi = \frac{1}{m}\sum_{i=1}^{m} 1\{y^{(i)} = 1\}$$

$$\mu_0 = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}}$$

$$\mu_1 = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}$$

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T$$

GDA visual classification

- if common Σ, the two gaussians are identical except for the mean

- the separation is a line of equidistant points to the two means