Machine Learning

4771

Instructor: Tony Jebara

Topic 9

- Continuous Probability Models
- Gaussian Distribution
- Maximum Likelihood Gaussian
- Sampling from a Gaussian

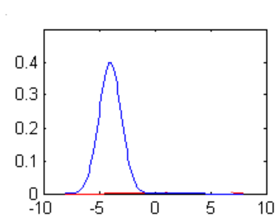
Continuous Probability Models

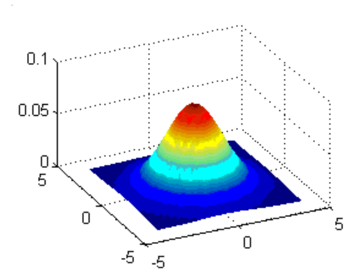
- Probabilities can have both discrete & continuous variables
- We will discuss:
  1) discrete probability tables

x=T	x=H
0.4	0.6

x=1	x=2	x=3	x=4	x=5	x=6
0.1	0.1	0.1	0.1	0.1	0.5

  2) continuous probability distributions



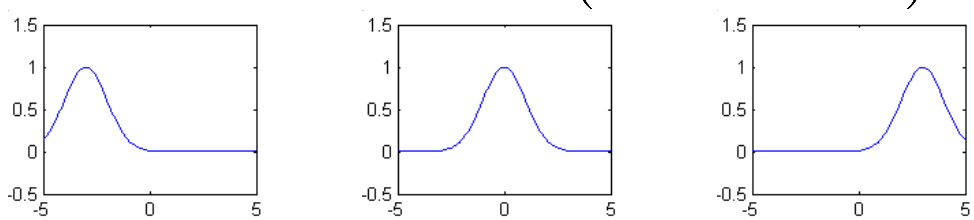


- Most popular continuous distribution = Gaussian

Gaussian Distribution

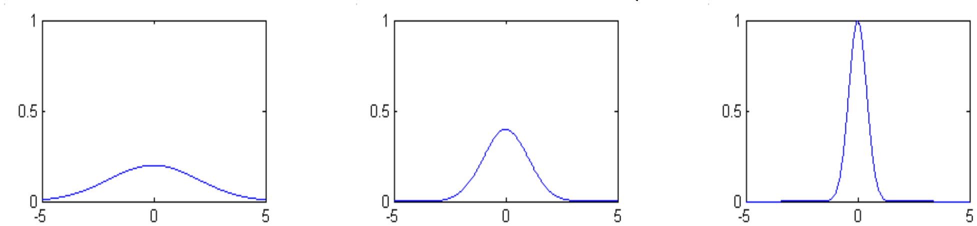
- Recall 1-dimensional Gaussian with mean parameter μ translates Gaussian left & right

$$p\left(x \mid \mu\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(x-\mu\right)^2\right)$$



- Can also have variance parameter σ^2 widens or narrows the Gaussian

$$p\left(x \mid \mu, \sigma\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right)$$



Note: $\int_{x=-\infty}^{\infty} p\left(x\right) dx = 1$

Multivariate Gaussian

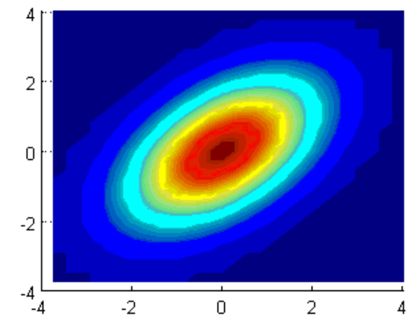
- Gaussian can extend to D-dimensions

- Gaussian mean parameter μ vector, it translates the bump
- Covariance matrix Σ stretches and rotates bump

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

$$\vec{x} \in \mathbb{R}^D, \vec{\mu} \in \mathbb{R}^D, \Sigma \in \mathbb{R}^{D\times D}$$

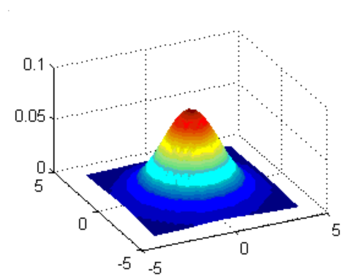
- Mean is any real vector
- Max and expectation = μ
- Variance parameter is now Σ matrix
- Covariance matrix is positive definite
- Covariance matrix is symmetric
- Need matrix inverse (inv)
- Need matrix determinant (det)
- Need matrix trace operator (trace)

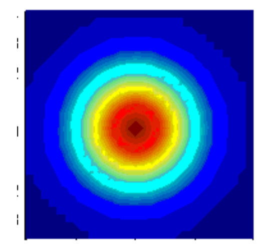


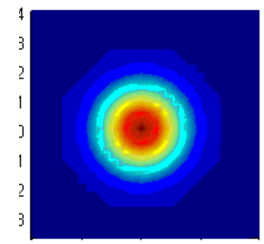
Multivariate Gaussian

•Spherical:

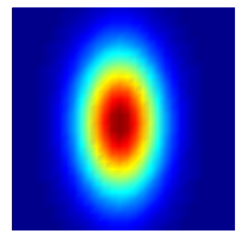
$$\Sigma = \sigma^2 I = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$

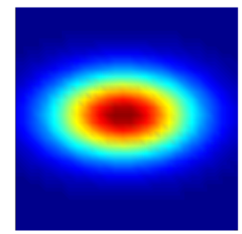






•Diagonal Covariance:
dimensions of x are independent
product of multiple 1d Gaussians





$$p\left(\vec{x} \mid \vec{\mu}, \Sigma\right) = \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi}\vec{\sigma}\left(d\right)} \exp\left(-\frac{\left(\vec{x}\left(d\right)-\vec{\mu}\left(d\right)\right)^2}{2\vec{\sigma}\left(d\right)^2}\right)$$

if the covariance :

$$\Sigma = \begin{bmatrix} \vec{\sigma}\left(1\right)^2 & 0 & 0 & 0 \\ 0 & \vec{\sigma}\left(2\right)^2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \vec{\sigma}\left(D\right)^2 \end{bmatrix}$$

Max Likelihood Gaussian

- Have IID samples as vectors i=1..N: $\mathcal{X} = \left\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N\right\}$
- How do we recover the mean and covariance parameters?
- Standard approach: Maximum Likelihood (IID)
- Maximize probability of data given model (likelihood)

Given the data,
Use the max-likelihood to
find the best parameters.

$$\Rightarrow p\left(\mathcal{X} \mid \theta\right) = p\left(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N \mid \theta\right)$$

$$= \prod_{i=1}^{N} p\left(\vec{x}_i \mid \vec{\mu}_i, \Sigma_i\right) \qquad \textit{independent Gaussian samples}$$

$$= \prod_{i=1}^{N} p\left(\vec{x}_i \mid \vec{\mu}, \Sigma\right) \qquad \textit{identically distributed}$$

- Instead, work with maximum of log-likelihood

$$L(u) = \sum_{i=1}^{N} \log p\left(\vec{x}_i \mid \vec{\mu}, \Sigma\right) = \sum_{i=1}^{N} \log \frac{1}{\left(2\pi\right)^{D/2}\sqrt{\left|\Sigma\right|}} \exp\left(-\frac{1}{2}\left(\vec{x}_i - \vec{\mu}\right)^T \Sigma^{-1}\left(\vec{x}_i - \vec{\mu}\right)\right)$$

$$L(u) = \sum_{i=1}^{N} -\frac{D}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| + (\log|\Sigma|) \cdot + (-\frac{1}{2}) \cdot (\vec{x}_i - \vec{u})^T \cdot \Sigma^{-1} \cdot (\vec{x}_i - \vec{u})$$

$$\frac{\partial L(u)}{\partial u} = \sum_{i=1}^{N} -\frac{1}{2} \cdot 2(\vec{x} - \vec{u}) \cdot \Sigma^{-1} = \sum_{i=1}^{N} 2(\vec{x} - \vec{u}) \cdot \Sigma^{-1} = 0$$

$$\sum_{i=1}^{N} (\vec{x}_i - \vec{u}) = 0$$

$$\vec{u} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Tony Jebara, Columbia University

Max Likelihood Gaussian : For u.

D: dimensionality

- Max over µ $\quad \frac{\partial}{\partial \mu}\left(\sum_{i=1}^{N} \log \frac{1}{(2\pi)^{D/2}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}\left(\vec{x}_i - \vec{\mu}\right)^T \Sigma^{-1}\left(\vec{x}_i - \vec{\mu}\right)\right)\right) = 0$

$$\frac{\partial}{\partial \mu}\left(\sum_{i=1}^{N} -\frac{D}{2}\log 2\pi - \frac{1}{2}\log\left|\Sigma\right| - \frac{1}{2}\left(\vec{x}_i - \vec{\mu}\right)^T \Sigma^{-1}\left(\vec{x}_i - \vec{\mu}\right)\right) = 0$$

Make a data Sheet

$$\boxed{\frac{\partial \vec{x}^T\vec{x}}{\partial \vec{x}} = 2\vec{x}^T}$$

$$\sum_{i=1}^{N}\left(\vec{x}_i - \vec{\mu}\right)^T \Sigma^{-1} = \vec{0}$$

see Jordan Ch. 12, get sample mean... $\quad \vec{\mu} = \frac{1}{N}\sum_{i=1}^{N}\vec{x}_i$

- For Σ need Trace operator: and several properties:

$$\left.\begin{aligned} &tr\left(A\right) = tr\left(A^T\right) = \sum_{d=1}^{D} A\left(d,d\right) \\ &tr\left(AB\right) = tr\left(BA\right) \\ &tr\left(BAB^{-1}\right) = tr\left(A\right) \\ &tr\left(\vec{x}\vec{x}^T A\right) = tr\left(\vec{x}^T A\vec{x}\right) = \vec{x}^T A\vec{x} \end{aligned}\right\}$$

Make a Data Sheet

Max Likelihood Gaussian ⇒ For covariance.

- Likelihood rewritten in trace notation:

$$l = \sum_{i=1}^{N} -\frac{D}{2}\log 2\pi - \frac{1}{2}\log\left|\Sigma\right| - \frac{1}{2}\left(\vec{x}_i - \vec{\mu}\right)^T \Sigma^{-1}\left(\vec{x}_i - \vec{\mu}\right)$$

$$= -\frac{ND}{2}\log 2\pi + \frac{N}{2}\log\left|\Sigma^{-1}\right| - \frac{1}{2}\sum_{i=1}^{N} tr\left[\left(\vec{x}_i - \vec{\mu}\right)^T \Sigma^{-1}\left(\vec{x}_i - \vec{\mu}\right)\right]$$

$$= -\frac{ND}{2}\log 2\pi + \frac{N}{2}\log\left|\Sigma^{-1}\right| - \frac{1}{2}\sum_{i=1}^{N} tr\left[\left(\vec{x}_i - \vec{\mu}\right)\left(\vec{x}_i - \vec{\mu}\right)^T \Sigma^{-1}\right]$$

$tr(\vec{x}\vec{x}^T A) = tr(\vec{x}^T A \vec{x})$

$$= -\frac{ND}{2}\log 2\pi + \frac{N}{2}\log\left|A\right| - \frac{1}{2}\sum_{i=1}^{N} tr\left[\left(\vec{x}_i - \vec{\mu}\right)\left(\vec{x}_i - \vec{\mu}\right)^T A\right]$$

- Max over $A=\Sigma^{-1}$ use properties:

$$\frac{\partial \log\left|A\right|}{\partial A} = \left(A^{-1}\right)^T \qquad \frac{\partial tr\left[BA\right]}{\partial A} = B^T$$

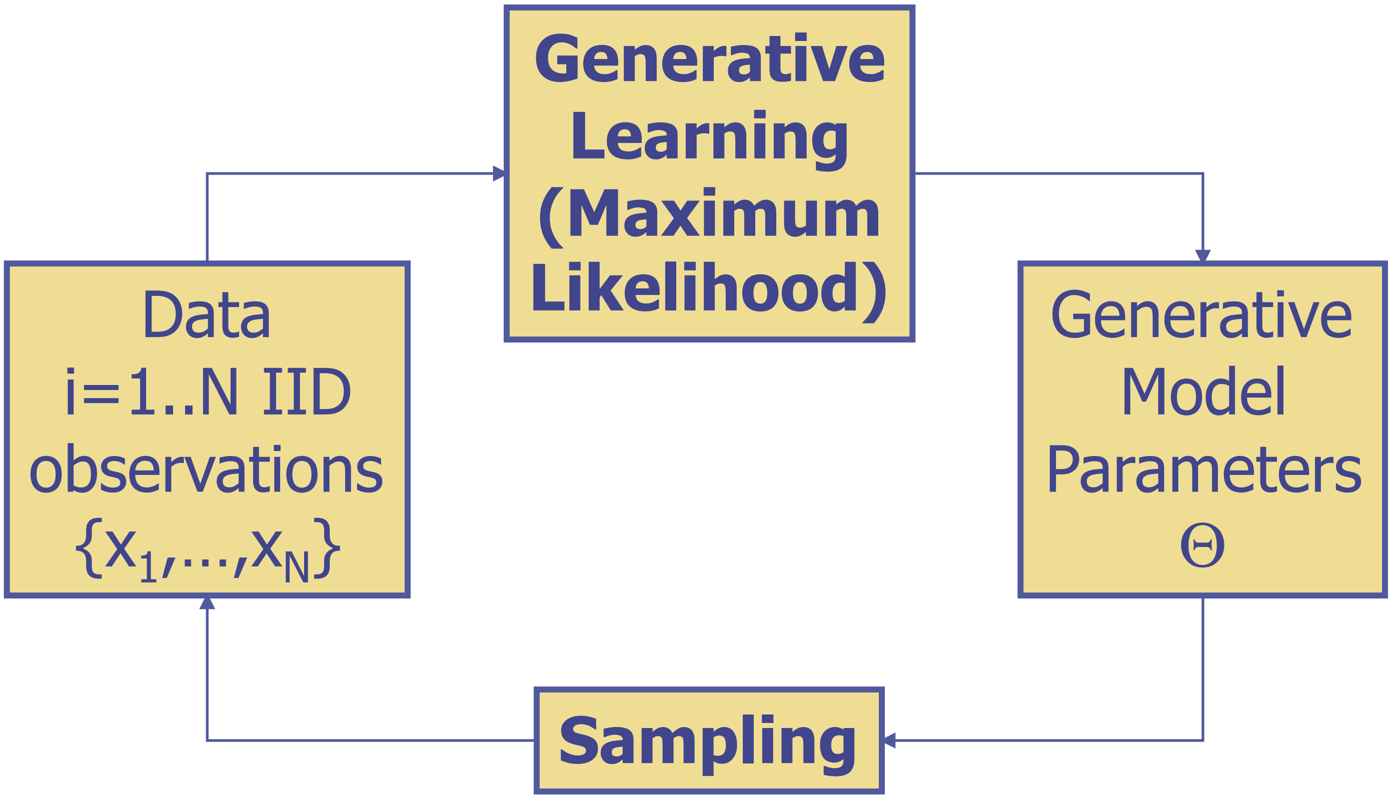
$$\frac{\partial l}{\partial A} = -0 + \frac{N}{2}\left(A^{-1}\right)^T - \frac{1}{2}\sum_{i=1}^{N}\left[\left(\vec{x}_i - \vec{\mu}\right)\left(\vec{x}_i - \vec{\mu}\right)^T\right]^T$$

$$= \frac{N}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{N}\left(\vec{x}_i - \vec{\mu}\right)\left(\vec{x}_i - \vec{\mu}\right)^T$$

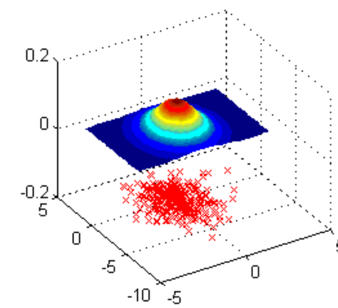
- Get sample covariance:

$$\frac{\partial l}{\partial A} = 0 \rightarrow \Sigma = \frac{1}{N}\sum_{i=1}^{N}\left(\vec{x}_i - \vec{\mu}\right)\left(\vec{x}_i - \vec{\mu}\right)^T$$

Sampling & Max Likelihood

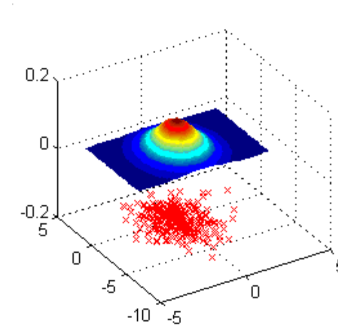


Sampling from a Gaussian



•Fit Gaussian to data, how is this Generative?

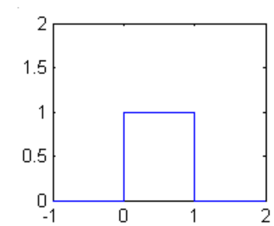
Sampling from a Gaussian



- Fit Gaussian to data, how is this Generative?
- Sampling! Generating discrete data easy:

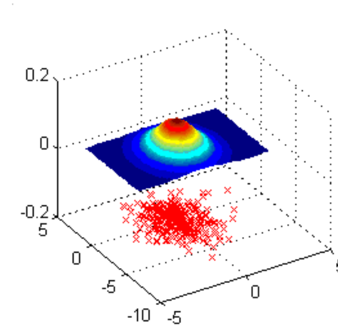
0.73	0.1	0.17

- Assume we can do uniform sampling:
  i.e. rand between (0,1)
  if 0.00 <= rand < 0.73 get A
  if 0.73 <= rand < 0.83 get B
  if 0.83 <= rand < 1.00 get C



- What are we doing?

Sampling from a Gaussian

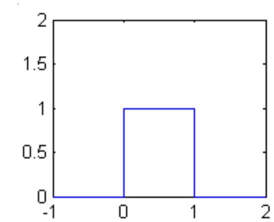


- Fit Gaussian to data, how is this Generative?
- Sampling! Generating discrete data easy:

0.73	0.1	0.17

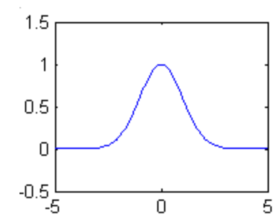
- Assume we can do uniform sampling:
  i.e. rand between (0,1)
  if 0.00 <= rand < 0.73 get A
  if 0.73 <= rand < 0.83 get B
  if 0.83 <= rand < 1.00 get C

For uniform Distribution.

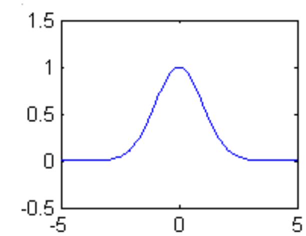


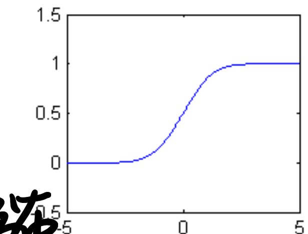
0.73	0.83	1.00

- What are we doing?
  Sum up the Probability Density Function (PDF)
  to get Cumulative Density Function (CDF)
- For 1d Gaussian, Integrate Probability Density Function get Cumulative Density Function
  Integral is like summing many discrete bars



Sampling from a Gaussian





- Integrate 1d Gaussian to get CDF:

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}x^2\right)$$

1-dimension

$$F(x) = \int_{-\infty}^{x} p(t)\,dt = \tfrac{1}{2} erf\left(\tfrac{1}{\sqrt{2}}x\right) + \tfrac{1}{2}$$ ⇒ CDF function

from uniform to Gaussian

- If sample from uniform, get: $u \sim uniform(0,1)$ 均匀分布
- Compute mapping: $x = F^{-1}(u) = \sqrt{2}erfinv(2u-1)$
- This is a Gaussian sample: $x \sim N(x \mid 0,1)$

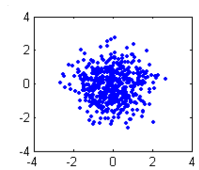
↓

D-Dimension

concatenate samples from 1-D samples

- For D-dimensional Gaussian N(z|0,I) concatenate samples:

$$\vec{x} = \left[\vec{x}(1) \ldots \vec{x}(D)\right]^T \sim p(\vec{x} \mid 0, I) = \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}\vec{x}(d)^2\right)$$

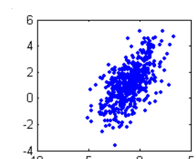


↓

Different μ, Σ.

- For N(z|μ,Σ), add mean & multiply by root cov

$$\vec{z} = \Sigma^{1/2}\vec{x} + \vec{\mu} \sim p(\vec{z} \mid \vec{\mu}, \Sigma)$$



- Example code: gendata.m