Machine Learning

4771

Instructor: Tony Jebara

Topic 9

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

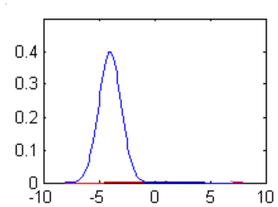
Continuous Probability Models

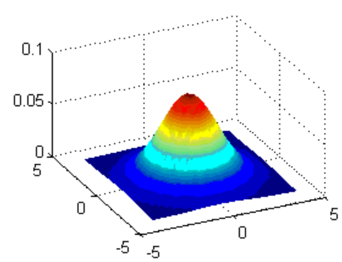
- Probabilities can have both discrete & continuous variables
- We will discuss:
  1) discrete probability tables

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



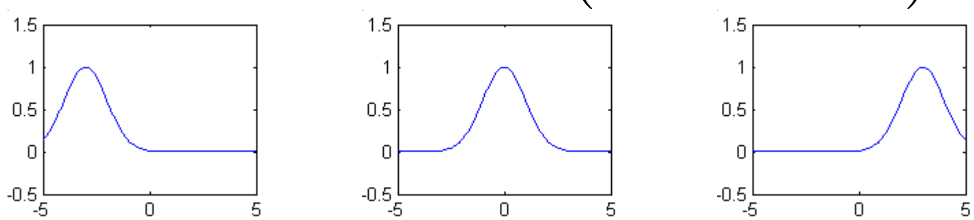


- Most popular continuous distribution = Gaussian

Gaussian Distribution

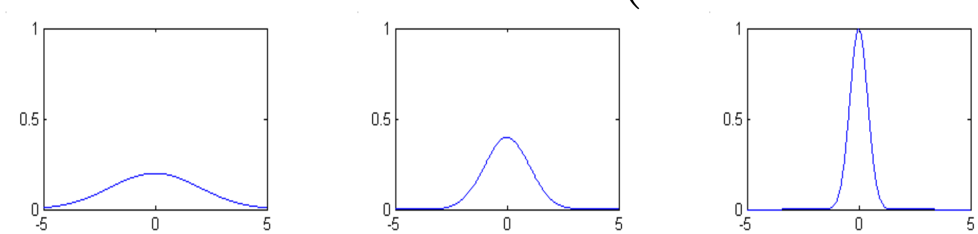
- Recall 1-dimensional Gaussian with mean parameter μ translates Gaussian left & right

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



- 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

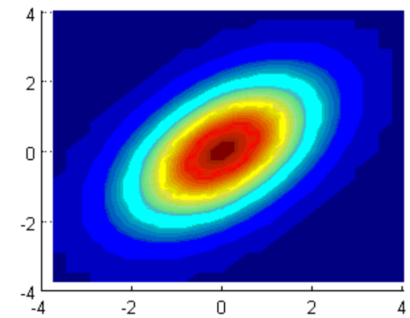
- Gaussian can extend to D-dimensions

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

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

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

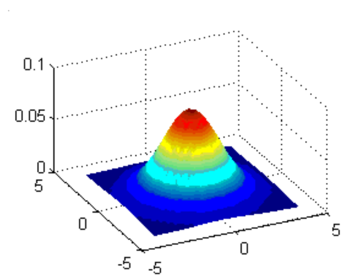
- Mean is any real vector
- Max and expectation = μ
- Variance parameter is now Σ matrix
- Covariance matrix is positive definite
- Covariance matrix is symmetric
- Need matrix inverse (inv)
- Need matrix determinant (det)
- Need matrix trace operator (trace)

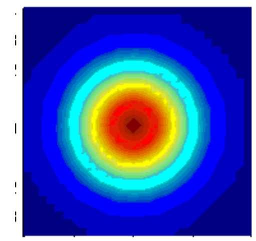


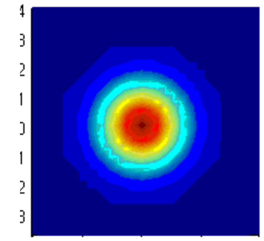
Multivariate Gaussian

- Spherical:

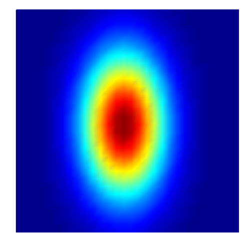
$$\Sigma = \sigma^2 I = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$

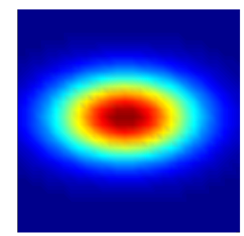






- Diagonal Covariance:
  dimensions of x are independent
  product of multiple 1d Gaussians





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

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

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)

$$\begin{aligned} p\left(\mathcal{X} \mid \theta\right) &= p\left(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N \mid \theta\right) \\ &= \prod_{i=1}^{N} p\left(\vec{x}_i \mid \vec{\mu}_i, \Sigma_i\right) \quad \textit{independent Gaussian samples} \\ &= \prod_{i=1}^{N} p\left(\vec{x}_i \mid \vec{\mu}, \Sigma\right) \quad \textit{identically distributed} \end{aligned}$$

- Instead, work with maximum of log-likelihood

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

Max Likelihood Gaussian

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

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

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

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

see Jordan Ch. 12, get sample mean...

$$\vec{\mu}=\tfrac{1}{N}\sum_{i=1}^{N}\vec{x}_i$$

- For Σ need Trace operator: $$tr\left(A\right)=tr\left(A^T\right)=\sum\nolimits_{d=1}^{D}A\left(d,d\right)$$

and several properties:

$$tr\left(AB\right)=tr\left(BA\right)$$

$$tr\left(BAB^{-1}\right)=tr\left(A\right)$$

$$tr\left(\vec{x}\vec{x}^TA\right)=tr\left(\vec{x}^TA\vec{x}\right)=\vec{x}^TA\vec{x}$$

Max Likelihood Gaussian

- Likelihood rewritten in trace notation:

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

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

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

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

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

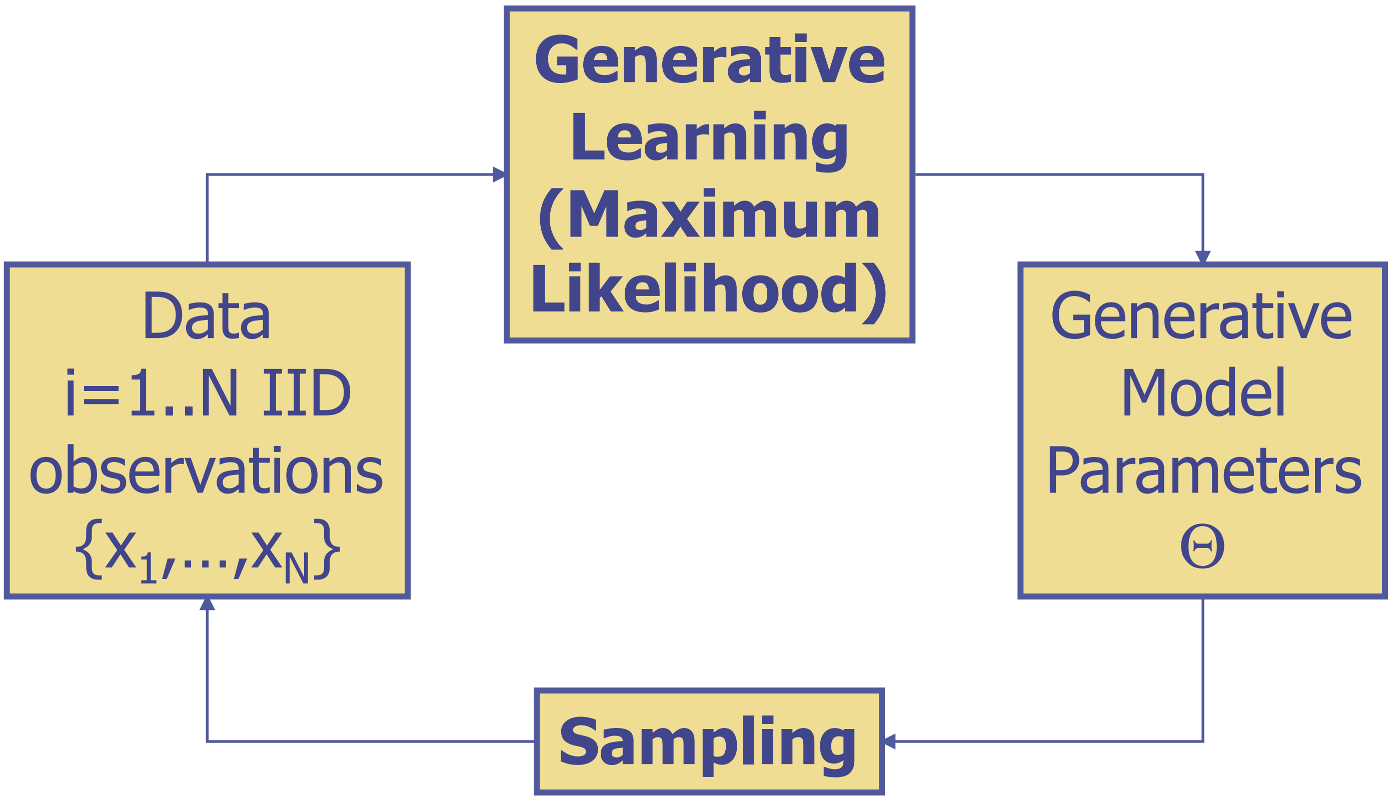
- Max over $A=\Sigma^{-1}$ use properties:

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

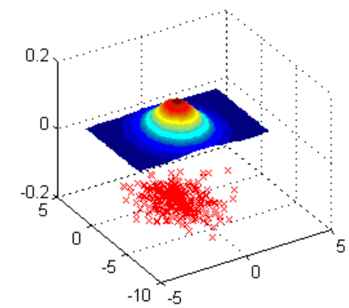
$$= \tfrac{N}{2}\Sigma - \tfrac{1}{2}\sum_{i=1}^{N}\left(\vec{x}_i - \vec{\mu}\right)\left(\vec{x}_i - \vec{\mu}\right)^T$$

- Get sample covariance: $\dfrac{\partial l}{\partial A} = 0 \;\rightarrow\; \Sigma = \tfrac{1}{N}\sum_{i=1}^{N}\left(\vec{x}_i - \vec{\mu}\right)\left(\vec{x}_i - \vec{\mu}\right)^T$

Sampling & Max Likelihood

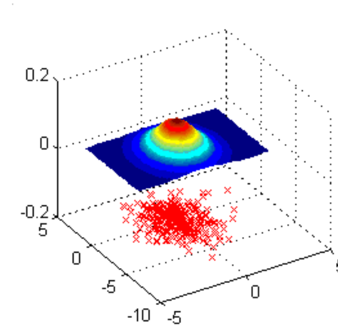


Sampling from a Gaussian



- Fit Gaussian to data, how is this Generative?

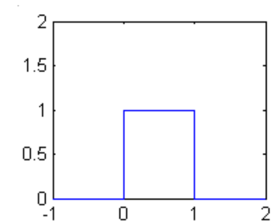
Sampling from a Gaussian



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

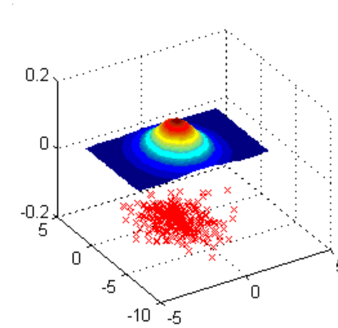
0.73	0.1	0.17

- 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?

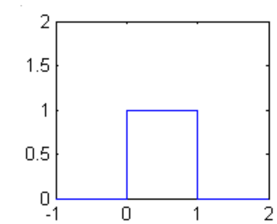
Sampling from a Gaussian



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

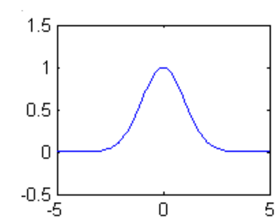
0.73	0.1	0.17

- 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

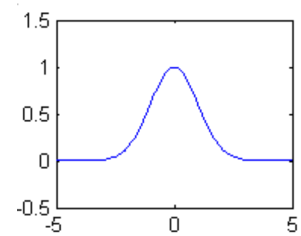


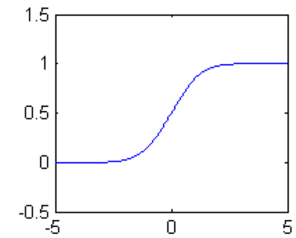
0.73	0.83	1.00

- 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





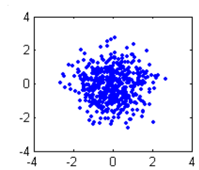
- Integrate 1d Gaussian to get CDF:

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

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

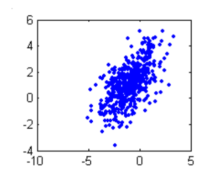
- If sample from uniform, get: $u \sim uniform\left(0,1\right)$
- Compute mapping: $x = F^{-1}\left(u\right) = \sqrt{2} erfinv\left(2u - 1\right)$
- This is a Gaussian sample: $x \sim N\left(x \mid 0,1\right)$
- For D-dimensional Gaussian N(z|0,I) concatenate samples:

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



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

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



- Example code: gendata.m