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 |

  2) continuous probability distributions

• Most popular continuous distribution = Gaussian
Gaussian Distribution

• Recall 1-dimensional Gaussian with mean parameter $\mu$ translates Gaussian left & right

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

• Can also have variance parameter $\sigma^2$ widens or narrows the Gaussian

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

Note: \( \int_{x=-\infty}^{\infty} p(x) dx = 1 \)
Multivariate Gaussian

- Gaussian can extend to D-dimensions
- Gaussian mean parameter $\mu$ vector, it translates the bump
- Covariance matrix $\Sigma$ stretches and rotates bump

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

- Mean is any real vector
- Max and expectation $= \mu$
- Variance parameter is now $\Sigma$ 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

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

\[
\Sigma = \begin{bmatrix}
\tilde{\sigma}(1)^2 & 0 & 0 & 0 \\
0 & \tilde{\sigma}(2)^2 & 0 & 0 \\
0 & 0 & \ddots & 0 \\
0 & 0 & 0 & \tilde{\sigma}(D)^2 
\end{bmatrix}
\]
Max Likelihood Gaussian

• Have IID samples as vectors i=1..N: \( \mathbf{x} = \{ \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \} \)

• How do we recover the mean and covariance parameters?

• Standard approach: Maximum Likelihood (IID)

• Maximize probability of data given model (likelihood)

\[
p(\mathbf{x} | \theta) = p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N | \theta) \\
= \prod_{i=1}^{N} p(\mathbf{x}_i | \mu_i, \Sigma_i) \quad \text{independent Gaussian samples} \\
= \prod_{i=1}^{N} p(\mathbf{x}_i | \mu, \Sigma) \quad \text{identically distributed}
\]

• Instead, work with maximum of log-likelihood

\[
\sum_{i=1}^{N} \log p(\mathbf{x}_i | \mu, \Sigma) = \sum_{i=1}^{N} \log \frac{1}{\sqrt{(2\pi)^{D/2}} \sqrt{\Sigma}} \exp \left( -\frac{1}{2} (\mathbf{x}_i - \mu)^T \Sigma^{-1} (\mathbf{x}_i - \mu) \right)
\]
Max Likelihood Gaussian

• Max over $\mu$

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

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

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

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

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

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

• For $\Sigma$ need Trace operator:

$$tr(A) = tr(A^T) = \sum_{d=1}^{D} A(d,d)$$

$$tr(AB) = tr(BA)$$

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

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

and several properties:
Max Likelihood Gaussian

- Likelihood rewritten in trace notation:
  \[ l = \sum_{i=1}^{N} -\frac{D}{2} \log 2\pi - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (\mathbf{x}_i - \mathbf{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \mathbf{\mu}) \]
  \[ = -\frac{ND}{2} \log 2\pi + \frac{N}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^{N} tr \left[ (\mathbf{x}_i - \mathbf{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \mathbf{\mu}) \right] \]
  \[ = -\frac{ND}{2} \log 2\pi + \frac{N}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^{N} tr \left[ (\mathbf{x}_i - \mathbf{\mu})(\mathbf{x}_i - \mathbf{\mu})^T \Sigma^{-1} \right] \]

- Max over \( A = \Sigma^{-1} \)
  - use properties:
    \[ \frac{\partial \log |A|}{\partial A} = (A^{-1})^T \]
    \[ \frac{\partial tr[BA]}{\partial A} = B^T \]
    \[ \frac{\partial}{\partial A} \left[ \sum_{i=1}^{N} (\mathbf{x}_i - \mathbf{\mu})(\mathbf{x}_i - \mathbf{\mu})^T \right] = \frac{N}{2} \sum_{i=1}^{N} (\mathbf{x}_i - \mathbf{\mu})(\mathbf{x}_i - \mathbf{\mu})^T \]
    \[ \frac{\partial l}{\partial A} = -0 + \frac{N}{2} (A^{-1})^T - \frac{1}{2} \sum_{i=1}^{N} \left[ (\mathbf{x}_i - \mathbf{\mu})(\mathbf{x}_i - \mathbf{\mu})^T \right] \]
    \[ = \frac{N}{2} \sum_{i=1}^{N} (\mathbf{x}_i - \mathbf{\mu})(\mathbf{x}_i - \mathbf{\mu})^T \]
    \[ \frac{\partial l}{\partial A} = 0 \rightarrow \Sigma = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \mathbf{\mu})(\mathbf{x}_i - \mathbf{\mu})^T \]
Sampling & Max Likelihood

Data i=1..N IID observations \( \{x_1, \ldots, x_N\} \)

Generative Learning (Maximum Likelihood)

Generative Model Parameters \( \Theta \)

Sampling
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:  

• 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:
  • 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?
    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:
  \[ p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} x^2\right) \]
  \[ F(x) = \int_{-\infty}^{x} p(t) \, dt = \frac{1}{2} \text{erf}\left(\frac{1}{\sqrt{2}} x\right) + \frac{1}{2} \]
- If sample from uniform, get: \( u \sim \text{uniform}(0, 1) \)
- Compute mapping:
  \[ x = F^{-1}(u) = \sqrt{2} \text{erfinv}(2u - 1) \]
- This is a Gaussian sample:
  \( x \sim N(x \mid 0, 1) \)

- For D-dimensional Gaussian \( N(z \mid 0, I) \) concatenate samples:
  \[ \vec{x} = \begin{bmatrix} \vec{x}(1) \cdots \vec{x}(D) \end{bmatrix}^T \sim p(\vec{x} \mid 0, I) = \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \vec{x}(d)^2\right) \]
- For \( N(z \mid \mu, \Sigma) \), add mean & multiply by root cov
  \[ \tilde{z} = \Sigma^{1/2} \vec{x} + \vec{\mu} \sim p(\tilde{z} \mid \vec{\mu}, \Sigma) \]
- Example code: gendata.m