Cody homework emailed. Due Monday and Wednesday before class.
Email me if you cannot attend as that way you can look at homework.

Announcements

Monday

• Gaussian 2.3
• Homework
• Information theory 1.6
• Decision theory 1.5
• Gaussian 1.2

Today:

Grade last year (A+ 19, A 20, A- 13, B+ 7, S 1, W 1)

Today: Introduction to ML, Lecture 20:)

Podcast might work eventually.

Piazza to come

Hi!
Curve Fitting Re-visited, Bishop1.2.5

\[
(\epsilon^{\prime}(\lambda^{0}x)h|\psi) \mathcal{N} = (\epsilon^{\prime}(\lambda^{0}x|\psi) \mathcal{d}
\]

\[
(\lambda^{0}x)h
\]
Maximum Likelihood Bishop 1.2.5

\[ \frac{d}{d \theta} \ln L = 0 \]

\[ \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 = \frac{n}{\sigma^2} \]

\[ \ln L(\mu) = \ln \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \]

\[ t = \frac{\bar{x} \pm t_{\alpha/2, n-1} \cdot s / \sqrt{n}}{s / \sqrt{n}} \]

\[ t \sim t_{\alpha/2, n-1} \]

\[ n \sim N(0, \sigma^2) = N(0, 1.5 \cdot 1) \]
(1.64) \[ \text{Given estimates of } W \text{ and beta, we can predict} \]

\[ \cdot \left( \prod_{i=1}^{N} \mathcal{N}(w_{i}, \mathbf{x}, \cdot | x) \right) \mathcal{N} = \left( \prod_{i=1}^{N} \mathcal{N}(w_{i}, \mathbf{x}, \cdot | x) \right) d \]

(1.63) \[ (w_2) \sum_{i=1}^{N} \frac{z_{i}}{N} - \frac{c}{N} \sum_{i=1}^{N} \frac{z_{i}}{N} + \frac{u_{i} - (\mathbf{w}^{T} \mathbf{x}) \mathbf{h}}{N} = \left( \frac{c}{N} \right) (\mathbf{w}, \mathbf{x}, \cdot | x) d \]

(1.62) \[ (w_2) \sum_{i=1}^{N} \frac{z_{i}}{N} - \frac{c}{N} \sum_{i=1}^{N} \frac{z_{i}}{N} + \frac{u_{i} - (\mathbf{w}^{T} \mathbf{x}) \mathbf{h}}{N} = \left( \frac{c}{N} \right) (\mathbf{w}, \mathbf{x}, \cdot | x) d \]

Form

Gaussian distribution, given by (1.46), we obtain the log likelihood function in the

Gaussian distribution, given by (1.46), we obtain the log likelihood function in the

As we did in the case of the simple Gaussian distribution earlier, it is convenient to

Maximun Likelihood

\[ (1.61) \]
Predictive Distribution

\[
\mathcal{N} = \int_{\mathcal{M}} g' \mathcal{N}^\mathcal{M}(x|\eta) \, d\eta
\]
Determine MAP by minimizing regularized sum-of-squares error, $E(w)$. 

$$w = \arg \min_{w} \{ u^t - (w, ux) \} + \varepsilon \sum_{i=1}^{N} \| \theta \|^2 = (w) \sim \mathcal{G}$$

$$\begin{align*}
\ln p(x | \mu) d(\theta | \mu, \Sigma) & \propto (x, t, a, \theta, \mu, \Sigma) \\
\frac{1}{N} \sum_{i=1}^{N} e^{-\frac{1}{2} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)} & = (x | \mu) N = (x | \mu) d
\end{align*}$$

MAP: A step towards Bayes 1.2.5
Important quantity in

• coding theory
• statistical physics
• machine learning

\[ (x)d \log \left( \frac{(x)d}{\sum_{x} x} \right) = [x]H \]
Differential Entropy

For fixed bins of width $c$ along the real line, put bins of width $c$ along the real line. In which case differential entropy maximized when

$$\int x \log y \, dx \bigg|_{x} = \left\{ \left( \int x \, dx \right) \log \left( \int x \, dx \right) \right\} \frac{c}{1} = [x]H$$

$$\left( \frac{x \log y}{c} \right) \mathcal{N} = (x) d$$

For fixed $\mathcal{O}$ differential entropy maximized when

$$\int x \, dx \bigg|_{x} = \left\{ \left( \int x \, dx \right) \log \left( \int x \, dx \right) \right\} \frac{c}{1}$$
The Kullback-Leibler divergence

\[ (d \| b)_{KL} \neq (b \| d)_{KL} \]

0 \leq (b \| d)_{KL}

\{ (u_x)d \ u \} + (\theta | u_x)b \ u \} \sim \frac{\sum_{i=1}^{u} N_i}{N} \approx (b \| d)_{KL}

\text{true distribution, q is approximating distribution}

\text{The Kullback-Leibler Divergence}
Decision Theory

Inference step
Determine either $d(x, t)$ or $d(x, t')$.

Decision step
For given $x$, determine optimal $t$. 

Decision Theory
\[ \int_{\mathbb{R}^d} \mathbb{P}(x \in C^1 | \mathcal{G}) p(x) \, dx + \int_{\mathbb{R}^d} \mathbb{P}(x \in C^2 | \mathcal{G}) p(x) \, dx = \int_{\mathbb{R}^d} \mathbb{P}(x \in C^1 | \mathcal{G}) p(x) \, dx + \int_{\mathbb{R}^d} \mathbb{P}(x \in C^2 | \mathcal{G}) p(x) \, dx = (\text{mistake rate}) \]

Minimum Misclassification Rate
Mixtures of Gaussians (Bishop 2.3.9)

Single Gaussian

Mixture of two Gaussians

Old Faithful geyser:
The time between eruptions has a bimodal distribution, with the mean interval being either 65 or 91 minutes, and is dependent on the length of the prior eruption. Within a margin of error of ±10 minutes, Old Faithful will erupt either 65 minutes after an eruption lasting less than \( \frac{1}{2} \) minutes, or 91 minutes after an eruption lasting more than \( \frac{1}{2} \) minutes.
Mixtures of Gaussians (Bishop 2.3.9)

- Combine simple models into a complex model:

$$\sum_{k=1}^{K} \begin{cases} \mathcal{N}(\mathbf{x} | \mu_k, \Sigma_k) \end{cases} = (\mathbf{x})d$$

Components:
- Mixing coefficient
- $K = 3$

MIXTURES OF GAUSSIANS (BISHOP 2.3.9)
Mixtures of Gaussians (Bishop 2.3.9)
Mixtures of Gaussians (Bishop 2.3.9).

- Determining parameters $\mu$, $\sigma$, and $\pi$ and using maximum log likelihood.

\[
\left\{ \sum_{l=1}^{M} \pi_l \mathcal{N}(x \mid \mu_l, \sigma_l^2) \right\} \prod_{n=1}^{N} = \int \prod_{n=1}^{N} \mathcal{N}(x \mid \mu, \sigma^2) \, dx \\
\]

- Expectation maximization algorithm (Chapter 9).

Solution: use standard, iterative, numeric optimization methods or the log of a sum; no closed form maximum.
Homework
Parametric Distributions

Basic building blocks:

Need to determine given

Recall Curve Fitting

We focus on Gaussians!

\[
\mathcal{N}(\mathbf{x} | \mathbf{\mu}) \text{ or } (\mathbf{\mu} | \mathbf{x}) \int = (\mathbf{1}, \mathbf{x}, x | \mathbf{\mu}) d
\]

Recall Curve Fitting

Reparameterization:

\[
\left\{ N \mathbf{x}, \ldots, \mathbf{x} \right\} \underbrace{(\mathbf{\theta}) d}_{\text{or } \mathbf{\theta}} \text{ or } \underbrace{\mathbf{\theta} | \mathbf{x} d}_{\text{Given}}
\]

Need to determine blocks:

Basic building blocks:

Parametric Distributions
The Gaussian Distribution

\[ \left\{ (\eta - x) \frac{\varepsilon}{I} - \varepsilon_x \right\} \exp \frac{\varepsilon}{I} \cdot \frac{\varepsilon}{\sigma(z)} N \left( \varepsilon, \eta \mid x \right) = (\varepsilon, \eta \mid x) N \]

\[ \left\{ \frac{\varepsilon}{I} \frac{\varepsilon}{\sigma(z)} \right\} \exp \frac{\varepsilon}{I} \cdot \frac{\varepsilon}{\sigma(z)} N \left( \varepsilon, \eta \mid x \right) = (\varepsilon, \eta \mid x) N \]
Central Limit Theorem

- The distribution of the sum of $N$ i.i.d. random variables becomes increasingly Gaussian as $N$ grows.
- Example: $N$ uniform [0,1] random variables.
- Gaussian as $N$ grows.
- The distribution of the sum of $N$ i.i.d. random variables becomes increasingly Gaussian.
Geometry of the Multivariate Gaussian

\[(\mathbf{r} - \mathbf{x}) \perp \mathbf{n} = \mathbf{f} \]

\[\mathbf{f} \quad \mathbf{n} \quad \mathbf{x} \]

\[\mathbf{f} \mathbf{n} \mathbf{x} \]
Moments of the Multivariate Gaussian (2)

A Gaussian requires \( D*(D-1)/2 \) parameters. Often we use \( D+1 \) parameters.

\[
\mathbf{I} = \begin{bmatrix}
1 & 0 \\
0 & 1
\end{bmatrix}
\]

\[
\widehat{\mathbf{\Sigma}} = \mathbf{I} \cdot \text{cov} \left( \mathbf{x} \right)
\]

\[
\mathbf{\Sigma} + \mathbf{I} = \mathbf{I} \cdot \text{cov} \left( \mathbf{x} \right)
\]
\[
\begin{align*}
\left( q \in \mathcal{X}, v \in \mathcal{X} \right)^N &= \left( q x - q x \right) + v x \\
q x y \left( q x, v x \right) d \int &= (v x) d
\end{align*}
\]
Given i.i.d. data, the log likelihood function is given by

\[
\ell(\mathbf{X}) = \sum_{i=1}^{N} \log p(\mathbf{x}_i | \mathbf{X}) = \sum_{i=1}^{N} \log p(\mathbf{x}_i | \mathbf{X}) = \sum_{i=1}^{N} \log p(\mathbf{x}_i | \mathbf{X})
\]

Maximum Likelihood for the Gaussian (1)
Set the derivative of the log likelihood function to zero, and solve to obtain

\[ 0 = \left( \mathbf{r} - u \mathbf{x} \right)_I \mathbf{C} = \mathbf{C} \left( \mathbf{x} \right) \mathbf{C} \mathbf{r} \]

Similarly, set the derivative for the Gaussian maximum likelihood for the Gaussian distribution.

\[ \mathbf{r} \]