

Machine Vision

October 21, 2018

1 Lecture 1

Bernoulli Distribution

$$Pr(x) = \lambda^x (1 - \lambda)^{1-x}, \lambda \in [0, 1], x \in \{0, 1\}$$

$$Pr(x) = Bern_x[\lambda]$$

Beta Distribution

$$Pr(\lambda) = \frac{\Gamma[\alpha + \beta]}{\Gamma[\alpha]\Gamma[\beta]} \lambda^{\alpha-1} (1 - \lambda)^{\beta-1}, \alpha, \beta > 0$$

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt = (z-1)!$$

$$E[\lambda] = \frac{\alpha}{\alpha + \beta}$$

$$B(p, q) = \frac{q-1}{p+q+1} B(p, q-1)$$

α, β decide the coin fact λ

Categorical Distribution

$$Pr(x = k) = \lambda_k$$

$$Pr(x) = Cat_x[\lambda]$$

Dirichlet Distribution

$$Pr(\lambda_1 \dots \lambda_K) = \frac{\Gamma[\sum_{k=1}^K \alpha_k]}{\prod_{k=1}^K \Gamma[\alpha_k]} \prod_{k=1}^K \lambda_k^{\alpha_k-1}$$

$$Pr(\lambda_1 \dots \lambda_K) = \text{Dir}_{\lambda_1 \dots \lambda_K}[\alpha_1, \alpha_2, \dots, \alpha_K]$$

Univariate Normal Distribution

$$Pr(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-0.5(x - \mu)^2/\sigma^2]$$
$$Pr(x) = \text{Norm}_x[\mu, \sigma^2]$$

Normal Inverse Gamma Distribution

$$Pr(\mu, \sigma^2) = \frac{\sqrt{\gamma}\beta^\alpha}{\sigma\sqrt{2\pi}\Gamma[\alpha]} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left[-\frac{2\beta + \gamma(\delta - \mu)^2}{2\sigma^2}\right]$$
$$Pr(\mu, \sigma^2) = \text{NormInvGam}_{\mu, \sigma^2}[\alpha, \beta, \gamma, \delta]$$

Multivariate Normal Distribution

$$Pr(\mathbf{x}) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp[-0.5(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)]$$

Normal Inverse Wishart

$$Pr(\mu, \Sigma) = \frac{\gamma^{D/2} |\Psi|^{\alpha/2} |\Sigma|^{-\frac{\alpha+D+2}{2}}}{(2\pi)^{D/2} 2^{\frac{\alpha D}{2}} \Gamma_D(\frac{\alpha}{2})} \exp\left\{-\frac{1}{2}(Tr(\Psi \Sigma^{-1})) + \gamma(\mu - \delta)^T \Sigma^{-1} (\mu - \delta)\right\}$$

Conjugate Distribution and Conjugate prior

Conjugate Distribution is between prior and posterior

Prior is the conjugate prior of the likelihood function.

2 Fitting model

maximum likelihood

Fitting

$$\begin{aligned}\hat{\theta} &= \underset{(\theta)}{\operatorname{argmax}} [Pr(\mathbf{x}_{1\dots I}|\theta)] \\ &= \underset{(\theta)}{\operatorname{argmax}} \left[\prod_{i=1}^I Pr(\mathbf{x}_i|\theta)\right]\end{aligned}$$

Predictive Density

Evaluate new data point \mathbf{x}^* under probability distribution $Pr(\mathbf{x}^*|\hat{\theta})$ with best parameter.

maximum a posteriori

Fitting

$$\begin{aligned}\hat{\theta} &= \operatorname{argmax}_{(\theta)} [Pr(\theta|\mathbf{x}_{1...I})] \\ &= \operatorname{argmax}_{(\theta)} \left[\frac{Pr(\mathbf{x}_{1...I}|\theta)Pr(\theta)}{Pr(\mathbf{x}_{1...I})} \right] \\ &= \operatorname{argmax}_{(\theta)} \left[\frac{\prod_{i=1}^I Pr(\mathbf{x}_i|\theta)Pr(\theta)}{Pr(\mathbf{x}_{1...I})} \right] \\ \hat{\theta} &= \operatorname{argmax}_{(\theta)} [Pr(\mathbf{x}_i|\theta)Pr(\theta)]\end{aligned}$$

Predictive

Evaluate new data point \mathbf{x}^* under probability distribution $Pr(\mathbf{x}^*|\hat{\theta})$ with best parameter.

bayesian approach

Fitting

$$Pr(\theta|\mathbf{x}_{1...I}) = \frac{(\prod_{i=1}^I Pr(\mathbf{x}_i|\theta))Pr(\theta)}{Pr(\mathbf{x}_{1...I})}$$

The difference between bayesian approach and MAP is that MAP takes the maximum value, while bayesian approach takes the distribution.

Predictive

$$Pr(\mathbf{x}^*|\mathbf{x}_{1...I}) = \int Pr(\mathbf{x}^*|\theta)Pr(\theta|\mathbf{x}_{1...I})d\theta$$

Confusion: the formula should be $\int Pr(\mathbf{x}^*|\theta, \mathbf{x}_{1...I})Pr(\theta|\mathbf{x}_{1...I})d\theta$. Given the θ , it considers $\mathbf{x}_{1...I}$ and x^* are independent.

Multivariate Normal Distribution

If $\mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_n$ are independent, the covariance matrix would be

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

Therefore, while $\mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_n$ are dependent, the covariance matrix could be decomposed into rotation matrix and diagonal:

$$\Sigma_{full} = \mathbf{R}^T \Sigma'_{diag} \mathbf{R}$$

Marginal Distribution

$$u_i = u_i$$

$$\Sigma_i = \Sigma_{ii}$$

Conditional Distribution

$$u_{i|j} = u_i + \Sigma_{ij} \Sigma_{jj}^{-1} (x_j - u_j)$$

$$\Sigma_{i|j} = \Sigma_{jj} - \Sigma_{ij}^T \Sigma_{ii}^{-1} \Sigma_{ij}$$

Product of two normals

$$\text{Norm}_{\mathbf{x}}[\mathbf{a}, \mathbf{A}] \text{Norm}_{\mathbf{x}}[\mathbf{b}, \mathbf{B}] = k \cdot \text{Norm}_{\mathbf{x}}[(\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}(\mathbf{A}^{-1}\mathbf{a} + \mathbf{B}^{-1}\mathbf{b}), (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}]$$

$$k = \text{Norm}_{\mathbf{a}}[\mathbf{b}, \mathbf{A} + \mathbf{B}]$$

change of variables

$$\text{Norm}_{\mathbf{x}}[\mathbf{A}\mathbf{y} + \mathbf{b}, \Sigma] = k \cdot \text{Norm}_{\mathbf{y}}[\mathbf{A}'\mathbf{x} + \mathbf{b}', \Sigma']$$

where

$$\mathbf{A}' = \Sigma' \mathbf{A}^T \Sigma^{-1}$$

$$\mathbf{b}' = -\Sigma' \mathbf{A}^T \Sigma^{-1} \mathbf{b}$$

$$\Sigma = (\mathbf{A}^T \Sigma^{-1} \mathbf{A})^{-1}$$

Learning and Inference

The observe measured data, \mathbf{x}

Draw inference from it about the state of world, \mathbf{w}

If \mathbf{w} is continuous, call this regression.

If \mathbf{w} is discrete, call this classification.

To compute the probability distribution $Pr(\mathbf{w}|\mathbf{x})$, we need: a model(related visual data \mathbf{x} and \mathbf{w} , the relationship depends on parameter θ), a learning algorithm(fits parameter θ from paired training examples $\mathbf{x}_i, \mathbf{w}_i$), an inference algorithm (use model to return $Pr(\mathbf{w}|\mathbf{x})$ given new observed data \mathbf{x})

Types of Model

1. Model contingency of the world on the data $Pr(\mathbf{w}|\mathbf{x})$ (Discriminative models)

1. Choose an appropriate form for $Pr(\mathbf{w})$
2. Make parameters a function of \mathbf{x}
3. Function takes parameters θ that define its shape.

Inference: evaluate $Pr(\mathbf{w}|\mathbf{x})$

2. Model joint occurrence of the world and data $Pr(\mathbf{x}, \mathbf{w})$ Generative models

1. Concatenate \mathbf{x} and \mathbf{w} to make $\mathbf{z} = [\mathbf{x}^T \mathbf{w}^T]$
2. Model of pdf of \mathbf{z}
3. Pdf takes parameter θ that define its shape

Inference: compute $Pr(\mathbf{w}|\mathbf{x})$ using Bayes rule.

$$Pr(\mathbf{w}|\mathbf{x}) = \frac{Pr(\mathbf{x}, \mathbf{w})}{Pr(\mathbf{x})} = \frac{Pr(\mathbf{x}, \mathbf{w})}{\int Pr(\mathbf{x}, \mathbf{w}) d\mathbf{w}}$$

3. Model contingency of data on the world $Pr(\mathbf{x}|\mathbf{w})$ (Generative models)

1. Choose an appropriate form for $Pr(\mathbf{x})$
2. Make parameters a function of \mathbf{w}
3. Function takes parameter θ that define its shape.

Inference: define prior $Pr(\mathbf{w})$ and then compute $Pr(\mathbf{w}|\mathbf{x})$ using Bayes' rule.

$$Pr(\mathbf{w}|\mathbf{x}) = \frac{Pr(\mathbf{x}|\mathbf{w})Pr(\mathbf{w})}{\int Pr(\mathbf{x}|\mathbf{w})Pr(\mathbf{w})d\mathbf{w}}$$

Bessel correction

$s^2 = (\frac{n}{n-1})s_n^2$ working this later.

Learning and inference

Mixture of Model

$$Pr(\mathbf{x}|\theta) = \sum_{k=1}^K Pr(\mathbf{x}, h = k|\theta)$$

Mixture of Gaussian

$$Pr(\mathbf{x}|\theta) = \sum_{k=1}^K \lambda_k \text{Norm}_{\mathbf{x}}[\mu_k, \Sigma_k]$$

Usually, the dimension would be smaller than sample.

Hidden variables

$$\begin{aligned} Pr(\mathbf{x}) &= \int Pr(\mathbf{x}, \mathbf{h}) d\mathbf{h} \\ Pr(\mathbf{x}|\theta) &= \int Pr(\mathbf{x}, \mathbf{h}|\theta) d\mathbf{h} \\ \hat{\theta} &= \text{argmax}_{\theta} \left[\sum_{i=1}^I \log \left[\int Pr(\mathbf{x}_i, \mathbf{h}_i|\theta) d\mathbf{h}_i \right] \right] \\ B[\{q_i(\mathbf{h}_i)\}, \theta] &= \sum_{i=1}^I \int q_i(\mathbf{h}_i) \log \left[\frac{Pr(\mathbf{x}, \mathbf{h}_i|\theta)}{q_i(\mathbf{h}_i)} \right] d\mathbf{h}_{1...I} \leq \sum_{i=1}^I \log \left[\int Pr(\mathbf{x}_i, \mathbf{h}_i|\theta) d\mathbf{h}_i \right] \end{aligned}$$

Lower bound

Because the log of sum is hard to derivate to 0.

According to Jensen's inequality when $f(x)$ is a convex function:

$$f(\mathbf{E}[\mathbf{X}]) \leq \mathbf{E}[f(\mathbf{X})]$$

For the concave function:

$$f(\mathbf{E}[\mathbf{X}]) \geq \mathbf{E}[f(\mathbf{X})]$$

Therefore the lower bound holds:

$$\begin{aligned}
\log(\mathbf{E} \left[\frac{Pr(\mathbf{x}, \mathbf{h}_i | \theta)}{q(\mathbf{h}_i)} \right]) &\geq \mathbf{E} \left[\log \left(\frac{Pr(\mathbf{x}, \mathbf{h}_i | \theta)}{q(\mathbf{h}_i)} \right) \right] \\
\log \left(\int \left[\frac{Pr(\mathbf{x}, \mathbf{h}_i | \theta)}{q(\mathbf{h}_i)} q(\mathbf{h}_i) \right] d\mathbf{h}_i \right) &\geq \int \left[q(\mathbf{h}_i) \log \left(\frac{Pr(\mathbf{x}, \mathbf{h}_i | \theta)}{q(\mathbf{h}_i)} \right) \right] d\mathbf{h}_i \\
\sum_{i=1}^I \log \left[\int Pr(\mathbf{x}_i, \mathbf{h}_i | \theta) d\mathbf{h}_i \right] &\geq \sum_{i=1}^I \int q_i(\mathbf{h}_i) \log \left[\frac{Pr(\mathbf{x}, \mathbf{h}_i | \theta)}{q_i(\mathbf{h}_i)} \right] d\mathbf{h}_{1 \dots I}
\end{aligned}$$

Where log function is the $f(\mathbf{X})$, and $q(\mathbf{h}_i)$ is $Pr(\mathbf{h}_i | \mathbf{x}_i, \theta^{[t]})$

E-Step

Maximize the bound w.r.t. distribution $q(\mathbf{h}_i)$

$$\hat{q}_i(\mathbf{h}_i) = Pr(\mathbf{h}_i | \mathbf{x}_i, \theta^{[t]}) = \frac{Pr(\mathbf{x}_i | \mathbf{h}_i, \theta^{[t]}) Pr(\mathbf{h}_i | \theta^{[t]})}{Pr(\mathbf{x}_i)}$$

M-Step

Maximize bound w.r.t parameter θ

$$\hat{\theta}^{[t+1]} = \operatorname{argmax}_{\theta} \left[\sum_{i=1}^I \int \hat{q}_i(\mathbf{h}_i) \log[Pr(\mathbf{x}_i, \mathbf{h}_i | \theta)] d\mathbf{h}_i \right]$$

E-step of MoG

$$\begin{aligned}
Pr(h_i = k | \mathbf{x}_i, \theta^{[t]}) &= \frac{Pr(\mathbf{x}_i | h_i = k, \theta^{[t]}) Pr(h_i = k, \theta^{[t]})}{\sum_{j=1}^K Pr(\mathbf{x}_i | h_i = j, \theta^{[t]}) Pr(h_i = j, \theta^{[t]})} \\
&= \frac{\lambda_k \operatorname{Norm}_{\mathbf{x}_i}[\mu_k, \Sigma_k]}{\sum_{j=1}^K \lambda_j \operatorname{Norm}_{\mathbf{x}_i}[\mu_j, \Sigma_j]} \\
&= r_{i,k}
\end{aligned}$$

M-step of MoG

Take derivative, equal to zero and solve:

$$\begin{aligned}
\lambda_k^{[t+1]} &= \frac{\sum_{i=1}^I r_{i,k}}{\sum_{j=1}^K \sum_{i=1}^I r_{i,j}} \\
\mu_k^{[t+1]} &= \frac{\sum_{i=1}^I r_{i,k} \mathbf{x}_i}{\sum_{i=1}^I r_{i,k}} \\
\Sigma_k^{[t+1]} &= \frac{\sum_{i=1}^I r_{i,k} (\mathbf{x}_i - \mu_k^{[t+1]})(\mathbf{x}_i - \mu_k^{[t+1]})^T}{\sum_{i=1}^I r_{i,k}}
\end{aligned}$$

Student t-distribution

not willing to write, seems not important

Factor analysis

not willing to write, seems not important