20CS6037 Machine Learning

MLE, MAP, Bayesian Reasoning - Chapter 3 & 5 (Lecture 5: 9/9/14)

Lecturer: Anca Ralescu *Scribes: Khaldoon Ashouiliy, Kyungmook Park*

Section 1:
Conditional Independence

Section 2:
Transformation of Random Variables

Section 3:
General Transformations

Section 4:
Monte Carlo Approximation

Section 5:
Entropy

Section 6:
Mutual Information

Section 1:
Conditional Independence

X,Y r.v. $X \perp Y$: X and Y are independent

Def
$X \perp Y \Leftrightarrow P(X,Y) = P(X)P(Y)$
This really means $\{w \in S \mid X(w) = a\}$, $\{w \in S \mid Y(w) = b\}$
are independent event for all $a \in$ Range(x); $b \in$ Range(Y)

Notation
$X \perp Y \mid 2$: X and Y are ***Conditionally Independent*** **(CI)** given 2

$P(X \perp Y \mid 2) \Leftrightarrow P(X, Y \mid 2) = P(X \mid 2)P(Y \mid 2)$

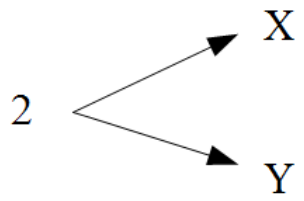

Figure 1: X and Y are CI given 2

P(X, Y) : Joint Distribution of X and Y
$\{w \in S \mid X(w) = a\}$, $\{w \in S \mid Y(w) = b\}$
Can extend to $P(X_1, \ldots, X_D)$: CDF, PDF/PMF

$$\begin{aligned} COV(X, Y) &\triangleq E[(X-E(X))(Y-E(Y))] \\ &= E[XY - XE(Y) - YE(X) + E(X)E(Y)] \\ &= E(XY) = E(X)E(Y) \end{aligned}$$

For a vector X = (X_1, X_2, X_3)

$$Cov[X] = \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & Cov(X_1, X_3) \\ Cov(X_2, X_1) & Var(X_2) & Cov(X_2, X_3) \\ Cov(X_3, X_1) & Cov(X_3, X_2) & Var(X_3) \end{bmatrix}$$

Cov $\in (0, \infty)$

Correlation $P(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \in$ [-1,1]

X, Y independent $\Rightarrow$ Cov(X,Y) = 0 $\Rightarrow$ $P(X,Y) = 0$ (Uncorrelated)

Independence $\rightleftharpoons$ Uncorrelated

ex)
X $\sim$ u(-1,1) ; Y = X^2 $\Rightarrow$ X, Y are dependent
$P(X,Y) = 0$

E(X) = $\frac{-1+1}{2} = 0$ Var(X) = $\frac{(1-(-1)^2}{12} = \frac{4}{12} = \frac{1}{3}$

$$\begin{aligned} E(Y) &= \int_{-1}^{1} x^2 f(x) dx \\ &= \int_{-1}^{1} x^2 (\tfrac{1}{2}) dx = \tfrac{1}{2} \tfrac{x^4}{4} |_{-1}^{1} = \tfrac{2}{6} = \tfrac{1}{3} \end{aligned}$$

E(XY) = $\int_{-1}^{1} x^3 \frac{1}{2} dx = \frac{1}{2} \frac{x^4}{4} |_{-1}^{1} = \frac{1}{2} \cdot 0$

Cov(X,Y) = E(XY) - E(X)E(Y) = 0 - 0 $\cdot \frac{1}{3}$ = 0 - 0 = 0

Section 2:
Transformation of Random Variables

X $\sim$ P(X) p:pdf
Y = f(X) What is the distribution of Y?
$P_Y(Y \leq y) = P_Y(f(X) \leq y) = \underline{P}_X(X \leq f^{-1}(y)) = P(f^{-1}(y))$

ex) Y = aX+b $\Rightarrow$ f(x) = aX+b

$f^{-1}(y) = \frac{y-b}{a}$ X = -1 Y = b-a ; X = 1 Y = a+b

$P_Y(y) = P_X\ (\frac{y-b}{a})$

ex) X $\sim$ u(-1,1)

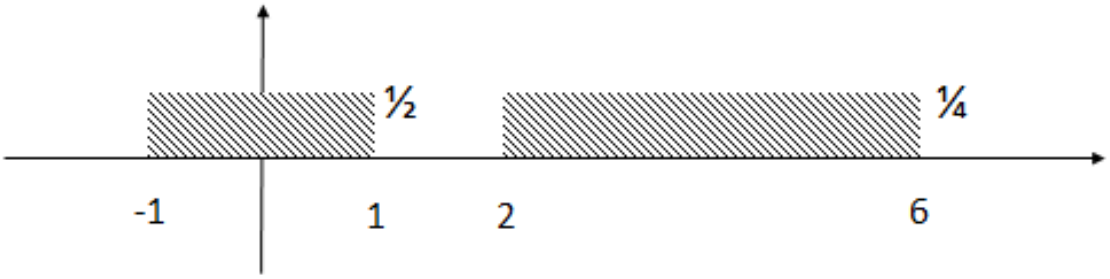

Figure 2: Result of Transformation

$$P_X(x) = \begin{cases} 0 & x \leq -1 \\ \frac{1}{2} & -1 < X < 1 \\ 0 & x \geq 1 \end{cases}$$

$$P_X(\frac{y-b}{a}) = \begin{cases} 0 & \frac{y-b}{a} \leq -1 \\ \frac{1}{b-a+a+b} & -1 < \frac{y-b}{a} < 1 \\ 0 & x \geq 1 \end{cases} = \begin{cases} 0 & \frac{y-b}{a} \leq -1 \\ \frac{1}{4} & -1 < \frac{y-b}{a} < 1 \\ 0 & x \geq 1 \end{cases}$$

$$\begin{cases} E(y) = & aE(x) + b \\ Var(y) = & a^2 E(x) \end{cases}$$

For multivariable case : X = (X_1, . . . , X_n), y=a^Tx+b
E(y) = y=a^TE(x)+b

Section 3:
General Transformations

Discrete Case X ∼ u(1, . . . , 4)
P(x=i) = $\frac{1}{4}$ i=1,2,3,4

$$Y = \begin{cases} 1 & if\ X\ is\ even \\ 0 & if\ X\ is\ odd \end{cases}$$

P(y = 1) $= P_x$ (x is even)
$\sum P_x(X = k) = P(X = 2) + P(X = 4) = \frac{2}{4} = \frac{1}{2}$

k ∈ {1, 2, 3, 4}
k is even

P(y=0) = . . . = $\frac{1}{2}$

Continuous Case
X ∼ P_x)(x) :pdf
Y = f(X) ⇒ P_Y(y) = ?

Use cdf P_Y(Y≤y) = P_Y(f(X)≤y) = P_X(X≤f^{-1}(y)) = P_X(f^{-1}(y))

Provided that f is invertible
Then P_Y(y) the pdf of Y is obtained by taking the derivative of cdf of Y

P_Y(y)=$\frac{d}{dy}P_Y$(Y≤y)=$\frac{d}{dy}P_X$(f^{-1}(y))=$\frac{d}{dx}P_X$(x)·$\frac{dx}{dy}$
Where x=f^{-1}(y) ignore sign of $\frac{dx}{dy}$
⇒ P_Y(y)=P_X(x)[$\frac{dx}{dy}$] ; For multivariable case J = $(\frac{\sigma y_i}{\sigma x_i})_{i,j}$
[det J]

example

X = (X_1,X_2);
Y = (γ,θ) : $X_1 = \gamma$Cosθ; $X_2 = \gamma$Sinθ

$$J = \begin{pmatrix} \frac{\sigma x_1}{\sigma\gamma} & \frac{\sigma x_1}{\sigma\theta} \\ \frac{\sigma x_2}{\sigma\gamma} & \frac{\sigma x_2}{\sigma\theta} \end{pmatrix} = \begin{pmatrix} Cos\theta & -\gamma Sin\theta \\ Sin\theta & \gamma Cos\theta \end{pmatrix}; det(J) = \gamma Cos^2\theta + \gamma Sin^2\theta = \gamma$$

|det(J)|=|γ| P_Y(y)=P_X(x)|det(J)|=|γ|P_X(x)

Section 4:
Monte Carlo Approximation

X, f(X)

Draw samples from X $x_1, \ldots, x_5$ (observations)
Approximate f(X) by the empirical distribution
of f(X) on $x_1, \ldots, x_5$

$\bar{u}$: A = $\bar{u}\gamma^2 \Rightarrow \bar{u} = \frac{A}{\gamma^2} = Approximate$
A=$4\gamma^2(\frac{1}{5})\sum_{i=1}^{5} f(x_i, y_i) f(x, y) = I(x^2 + y^2 \leq \gamma^2)$

Section 5: Entropy

Information Theory Shannon 1949
X∼P P:pmf H(X):entropy P_k=P(X=k)
H(X)=-$\sum_k P_k \log P_k$

$$X = \begin{cases} 1 & \theta \\ 0 & 1-\theta \end{cases}$$

$H_\theta(\mathrm{X}) = [\theta \log \theta + (1-\theta)\log(1-\theta)]$

$Max_\theta H_\theta(\mathrm{X})$ occurs at $\theta = \frac{1}{2}$
$\theta = \frac{1}{2}$-$(\frac{1}{2}(-1) + \frac{1}{2}(-1)) = 1$

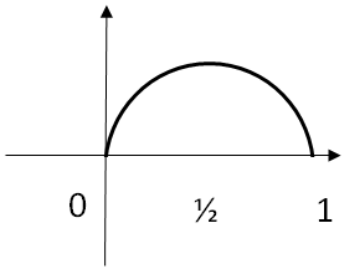

Figure 3: Entropy graph

Difference/Similarity between probability distributions.

KL Divergence

KL(p||q)$\triangleq \sum_{k=1}^{K} P_k \log \frac{P_k}{q_k} = \sum_k P_k \log P_k - \sum_k P_k \log q_k = -H(P) + H(P,q)$
H(P,q) is called the cross entropy.
Theorem 2.8.1 The Information Inequality
KL(P||q) ≥ 0

KL(P||q) = 0 $\Leftrightarrow$ P = q
The discrete distribution with maximum entropy is uniform distribution.

Section 6: Mutual Information

X, Y random variables
I(X;Y) = KL(P(X,Y) || P(X)P(Y)) = $\sum_x \sum_y P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \geq 0$
I(X;Y)=0 $\Leftrightarrow$P(x,y)=P(x)P(y) : independent
I(X;Y)=H(X)-H(X|Y) = H(Y) - H(Y|X) = $\sum_y P(y)H(X|Y=y)$
H(X|Y) is called the conditional entropy

PMI = Pointwise Mutual Information
PMI = $\log \frac{P(x,y)}{P(x)P(y)} = \log \frac{P(x|y)}{P(x)} = \log \frac{P(y|x)}{P(y)}$

For continuous, it is common to first discretize or quantize them by dividing the ranges of each variable into bins, and computing how many values fall in each histogram bin (Scott 1979)

MIC = Maximal Information Coefficient
m(X,Y) = $\frac{max_G I(X(G);Y(G))}{\log min(X,Y)}$
MIC= max(X,Y)
x,y:xy¡B
B is the bound on the number of bins.