Statistical Learning February 23, 2019
Sharif University of Technology CE 40-956
H.R. Rabiee

Time: 25 mins

Name: Std. Number:

Quiz 1 (Gaussian Processes)

Questions

1. Let $\{X(t); t \in \mathbb{R}\}$ be defined by $X(t) = tA$ for all $t \in \mathbb{R}$ where $A \sim N(0, 1)$. Show that $X(t)$ is a Gaussian process. Find its mean for each t and its covariance function.

2. Let $\{W(n); n \in \mathbb{Z}\}$ be a discrete Gaussian Process for which we have $W(n) \sim N(0, 1)$.
   (a) Find the mean and autocovariance function of this process. Is it *WSS*?
   (b) Let $\{S(n) = W(1) + W(2) + ... + W(n); n \geq 1\}$ be a new cumulative process.
   - Show that $S(n)$ is also a gaussian process.
   - Find it's mean and autocovariance function. Is $S(n)$ stationary in any sense?

Quiz-1, Solution

Quiz 1

(1)

برای هر تعداد انتخاب k از زمان های دلخواه $t_1, t_2, \dots, t_k$ متغیرهای $X_{t_1}, X_{t_2}, \dots, X_{t_k}$ برابر با متغیرهای تصادفی $At_1, At_2, \dots, At_k$ هستند.

توزیع توام این متغیرها را می توان طبق قاعده زنجیره ای به صورت زیر نوشت:

$$f(At_1, At_2, \dots, At_k) = f(At_1) \times f(At_2|At_1) \times \dots \times f(At_k|At_{k-1}, \dots, At_1)$$

$f(At_1)$ دارای فرم توزیع نرمال است. باقی عبارات هم به صورت مقادیر ثابت هستند. بنابراین توزیع مشترک $X_{t_1}, X_{t_2}, \dots, X_{t_k}$ فرم یک توزیع نرمال را دارد و بنا بر تعریف فرآیند حاصل یک فرآیند نرمال است.

- $E[tA] = tE[A] = 0$

$$E[X(t_1)X(t_2)] = E[A^2]t_1t_2 = t_1t_2$$

②

امید ریاضی این فرآیند برای هر n صفر است، و Auto Covariance آن برابر است با:

$$C_w(k,l) = E\left[w(k)\,w(l)\right] = \sigma^2 \delta_{k,l} \longrightarrow \delta_{k,l} = \begin{cases} 0 & k \neq l \\ 1 & k = l \end{cases}$$

در نتیجه با توجه به تعریف این فرآیند ایستا از نوع ضعیف است. (WSS است)

می‌توانیم بردار $(S(1), \dots, S(n))^T$ را به عنوان یک تبدیل خطی از بردار $(w(1), \dots, w(n))^T$ در نظر بگیریم. در این صورت $S_1, \dots, S_n$ یک مجموعه از متغیرهای مشترکاً گوسی با امید ریاضی صفر است. از آن‌جایی که این حرف برای هر $n \geq 1$ درست است، پس $\{S(n); n \geq 1\}$ یک فرآیند گوسی است.

$$E[S(n)] = E\left[w(1) + w(2) \dots, w(n)\right] = \sum_{i=1}^{n} E[w(i)] = 0$$

اگر $k \geq l$ برای Auto Covariance این فرآیند داریم:

$$C_s(k,l) = E\left[\sum_{i=1}^{k} w_i \sum_{j=1}^{l} w_j\right] = \sum_{i=1}^{k} E\left[w_i^2\right] = k\sigma^2$$

با استدلال مشابه حالت $k > l$ هم قابل محاسبه است. بنابراین برای حالت کلی داریم:

$$C_s(k,l) = \min[k,l]\,\sigma^2$$

همانطور که مشاهده می‌شود این فرآیند حتی از نوع ضعیف هم ایستا نیست. (WSS هم نیست)

Statistical Learning
Sharif University of Technology
H.R. Rabiee

March 2, 2019
CE 40-956

Name: Std. Number:

Quiz 2 (Gaussian Processes)

Questions

1. Let $x(t)$ be a normal process with zero expected value. If $x(t)$ is passed through a nonlinear system and we have $y(t) = x(t)^2$:

   (a) Show that
   $$S_y(\omega) = 2\pi R_x^2(0)\delta(\omega) + 2S_x(\omega) * S_x(\omega)$$
   ($S(\omega)$ is the spectral density of a process)

   (b) If $S_x(\omega)$ is an ideal low-pass filter, what does $S_y(\omega)$ look like?

2. Let $X(t) = R\cos(2\pi ft + \theta)$ where R is a Rayleigh rv and the rv θ is independent of R and uniformly distributed over the interval 0 to 2π.

   (a) Show that $E[X(t)] = 0$

   (b) Show that $E[X(t)X(t+\tau)] = \frac{1}{2}E[R^2]\cos(2\pi f\tau)$.

   (c) Show that $X(t)$; $t \in \mathbb{R}$ is a Gaussian process.

Quiz-2 Solution (Gaussian Processes)

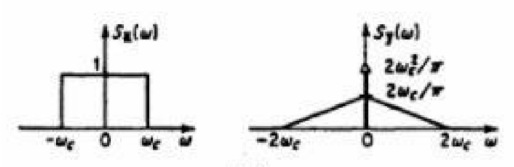
1. Solution:

   (a) According to the convolution theorem in frequency domain we have:

   $$R_y(\tau) = E[X^2(t+\tau)X^2(t)] = E[X^2(t+\tau)]E[X^2(t)] + 2E^2[X(t+\tau)X(t)] = R_x^2(0) + 2R_x^2(\tau)$$

   $$S_y(\omega) = 2\pi R_x^2(0)\delta(\omega) + 2S_x(\omega) * S_x(\omega)$$

   (b) $S_y(\omega)$ would look like this:



2. Solution:

   (a) This can be done by standard (and quite tedious) manipulations, but if we first look at $t = 0$ and condition on a sample value of R, we are simply looking at $\cos(\theta)$, and since θ is uniform over $[0, 2\pi)$, it seems almost obvious that the mean should be zero. To capture this intuition, note that $\cos(\theta) = -\cos(\theta + \pi)$. Since θ is uniform between 0 and 2π, $E[\cos(\theta)] = E[\cos(\theta + \pi)]$, so that $E[\cos(\theta)] = 0$. The same argument works for any t, so the result follows.

   (b) Since θ and R are independent, we have:

   $$E[X(t)X(t+\tau)] = E[R^2]E[\cos(2\pi ft + \theta)cos(2\pi f(t+\tau) + \theta)]$$

   $$= E[R^2]\frac{1}{2}E[\cos(4\pi ft + 2\pi f\tau + 2\theta) + \cos(2\pi f\tau)]$$

   $$= \frac{E[R^2]\cos(2\pi f\tau)}{2}$$

   (c) Let W_1, W_2 be iid normal Gaussian rv's. These can be expressed in polar coordinates as $W_1 = R\cos\theta$ and $W_2 = R\sin\theta$, where R is Rayleigh and θ is uniform. The rv $R\cos\theta$ is then $N(0,1)$. Similarly, $X(t)$ is a linear combination of W_1 and W_2 for each t, so each set $X(t_1), X(t_2), ...X(t_k)$ of rv's is jointly Gaussian. It follows that the process is Gaussian.

Statistical Machine Learning
Sharif University of Technology
Hamid R. Rabiee

March 7, 2020
CE 40-956

TH Quiz 2 (Gaussian Processes)

Due March 15, 2020 (11:59 pm)

1. Assume that $W(t)$ is a white Gaussian noise with autocorrelation function $\mathbf{R}_W(\tau) = \alpha\delta(\tau)$. $X(t)$ is obtained by passing $W(t)$ through an integrator:

$$X(t) = \int_0^t W(\tau)d\tau$$

   (a) Find $\mathbb{E}[X(t)]$ and $\mathbf{R}_X(t, t+\tau)$.

   (b) Show that $X(t)$ is nonstationary.

2. A process $W(t)$ is called Brownian if $W(0) = 0$ and for $\tau > 0, W(t+\tau) - W(t)$ is a Gaussian $\mathcal{N}(0, \sqrt{\alpha\tau})$ independent of $W(t')$ for all $t' \leq t$. Show that A Brownian motion process is a Gaussian process.

Quiz 2 (Gaussian Processes) Solutions

1 -a)

$$E[X(t)] = E\left[\int_0^t W(\tau)\, d\tau\right] = \int_0^t E[W(\tau)]\, d\tau = \int_0^t 0\ d\tau = 0$$

$$R_X(t, t+\tau) = E[X(t)X(t+\tau)] = E\left[\int_0^t \int_0^{t+\tau} W(u)W(v)\, dv\, du\right]$$

$$= \int_0^t \int_0^{t+\tau} E[W(u)W(v)]\, dv\, du$$

$$= \int_0^t \int_0^{t+\tau} \alpha\delta(u-v)\, dv\, du$$

Case 1: $\tau \geq 0$

$$R_X(t, t+\tau) = \int_0^t \alpha\, du = \alpha t$$

Case 2: $\tau < 0$

$$R_X(t, t+\tau) = \int_0^{t+\tau} \int_0^t \alpha\delta(u-v)\ du\, dv = \int_0^{t+\tau} \alpha\, dv = \alpha(t+\tau)$$

$$\Rightarrow R_X(t, t+\tau) = \alpha \min\{t, t+\tau\}$$

1-b)

Since $Var\big(X(t)\big) = \alpha.t$ is dependent on t, it cannot be stationary.

2)

Let $W = [W(t_1),\ W(t_2),\ \ldots, W(t_n)]^T$ denote a vector of samples of a Brownian motion process. To prove that W(t) is a Gaussian process, we must show that W is a Gaussian random vector. To do so, let

$$X = [X_1, X_2, \ldots, X_n]^T = [W(t_1), W(t_2) - W(t_1), \ldots, W(t_n) - W(t_{n-1})]^T$$

denote the vector of increments. By the definition of Brownian motion, $X_1, X_2, \ldots, X_n$is a sequence of independent Gaussian random variables. Thus, X is a Gaussian random vector. Finally,

$$W = \begin{bmatrix} X_1 \\ X_1 + X_2 \\ \dots \\ X_1 + X_2 + \cdots + X_n \end{bmatrix} = \begin{bmatrix} 1 & \cdots & \dots & 0 \\ 1 & 1 & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 1 & \cdots & 1 & 1 \end{bmatrix} . X$$

Since X is a Gaussian random vector and $W = AX$, where A is a rank n matrix, so W is also a Gaussian random vector.

Statistical Machine Learning
Sharif University of Technology
Hamid R. Rabiee

March 19, 2020
CE 40-956

TH Quiz 3 (Dirichlet Process - Chinese Restaurant Process)

Due April 7, 2020 (11:59 pm)

1. Let $G \sim DP(\alpha, G_0)$ be a dirichlet process.
   (a) If we increase/decrease α, predict how this change will affect the similarity between distributions G and G_0.
   (b) According to your answer to the last question, How will you choose α if distribution of your data consists of many/few clusters? Justify your answer.

2. Imagine a Chinese Restaurant Process running with the concentration parameter equal to α. If there are 10 customers in the restaurant and cluster (table) assignments are:

$$c_1, c_1, c_2, c_3, c_1, c_3, c_3, c_1, c_4, c_4$$

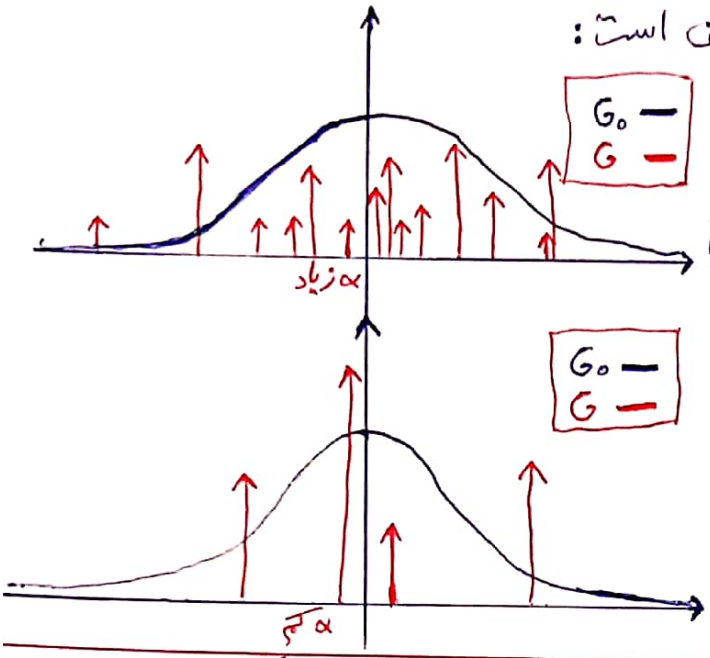
   (a) Compute the probability of this occurrence.
   (b) Change the order of cluster assignments such that the final clusters be equal to the last part. Then compute the probability of this occurrence again. Is CRP exchangeable?

3. Assume $G \sim DP(\alpha, G_0)$ is a Dirichlet process where G_0 is a probability measure defined over a set (Θ).
   (a) Find the mean of $G(A)$ for every $A \subset \Theta$.
   (b) Find the variance of $G(A)$ for every $A \subset \Theta$.

4. Assume $(\pi_1, \pi_2, \ldots, \pi_k) \sim Dirichlet(\alpha_1, \alpha_2, \ldots, \alpha_k)$. Show that:

$$(\pi_1 + \pi_2, \pi_3, \ldots, \pi_k) \sim Dirichlet(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_k).$$

۱ - a) موجودیت G_0 می‌تواند یک توزیع پیوسته (مثل گاوسی) باشد. اما همانطور که در شکل زیر مشاهده می‌کنید، موجودیت G مجموعه‌ای از توابع δ با دامنه‌های مختلف است:



اگر تعداد توابع δ را زیاد کنیم، توزیع G حاصل شباهت بیشتری با توزیع اولیه (G_0) خواهد داشت.

(دلیل این شباهت، انتخاب شدن مجدد توابع δ، طبق توزیع G_0 است)

با افزایش α، متد Stick Breaking، چوب را به تکه‌های کوچکتر (و در نتیجه پرتعداد‌تر) تقسیم می‌کند.

لذا α ی زیاد، نتیجه در شباهت G_0 و G داشته و با کاهش α، شباهت G_0 و G کمتر می‌شود.

---

۱ - b) در صورتی که توزیع مدنظر، تعداد خوشه‌های زیادی داشته باشد، باید مقدار α را زیاد در نظر بگیریم تا توزیع خروجی (G) با توزیع مدنظر (داده‌ها) تطابق بیشتری داشته باشد. (و بالعکس)

---

۲ - a)

ترتیب → c_1 c_1 c_2 c_3 c_1 c_3 c_3 c_1 c_4 c_4

$$P = \frac{\alpha}{\alpha} \times \frac{1}{\alpha+1} \times \frac{\alpha}{\alpha+2} \times \frac{\alpha}{\alpha+3} \times \frac{2}{\alpha+4} \times \frac{1}{\alpha+5} \times \frac{2}{\alpha+6} \times \frac{3}{\alpha+7} \times \frac{\alpha}{\alpha+8} \times \frac{1}{\alpha+9} =$$

$$= \frac{\alpha^4 \times 12}{\prod_{i=0}^{9} (\alpha+i)}$$

---

۲ - b)

ترتیب → c_1 c_1 c_1 c_1 c_2 c_3 c_3 c_3 c_4 c_4

$$P = \frac{\alpha}{\alpha} \times \frac{1}{\alpha+1} \times \frac{2}{\alpha+2} \times \frac{3}{\alpha+3} \times \frac{\alpha}{\alpha+4} \times \frac{\alpha}{\alpha+5} \times \frac{1}{\alpha+6} \times \frac{2}{\alpha+7} \times \frac{\alpha}{\alpha+8} \times \frac{1}{\alpha+9} =$$

$$= \frac{\alpha^4 \times 12}{\prod_{i=0}^{9} (\alpha+i)} = \text{همان حالت قبل}$$

توجه: حرج کسرها ثابت است. هر خوشه، با n عضو، جمله‌ی $\alpha \times (n-1)!$ را در صورت اضافه می‌کند.

چون با تغییر دادن ترتیب، این فرمول‌ها تغییری نمی‌کند، لذا exchangability در CRP صادق است.

۳- a) طبق تعاریف Consistency کولموگروف، اگر $G \sim DP(G_0, \alpha)$ و G_0 در فضای Θ تعریف شده باشد برای هر افراز $(A_1 \dots A_k)$ از فضای Θ، داریم:

$$\left(G(A_1), G(A_2), \dots, G(A_k)\right) \sim Dir\left(\alpha G_0(A_1), \alpha G_0(A_2), \dots, \alpha G_0(A_k)\right)$$

لذا اگر $k=2$ و $\left(A_1 \triangleq A \;,\; A_2 \triangleq A^c\right)$ باشد:

$$\left(G(A), G(A^c)\right) \sim Dir\left(\alpha G_0(A), \alpha G_0(A^c)\right)$$

$G(A^c) \to 1 - G(A)$

$\alpha G_0(A^c) \to \alpha\left(1 - G_0(A)\right)$

توزیع Dirichlet دو متغیره، همان توزیع بتا است.

لذا داریم:

$$G(A) \sim Beta\left(\underbrace{\alpha_0 G_0(A)}_{p}, \underbrace{\alpha_0\left(1 - G_0(A)\right)}_{q}\right)$$

طبق روابط توزیع بتا، داریم:

$$\mathbb{E}\left[G(A)\right] = \frac{p}{p+q} = \frac{\alpha_0 G_0(A)}{\alpha_0 G_0(A) + \alpha_0\left(1 - G_0(A)\right)} = G_0(A)$$

۳- b) باز هم طبق روابط توزیع بتا، داریم:

$$Var\left[G(A)\right] = \frac{pq}{(p+q)^2(p+q+1)} = \frac{\alpha_0^2 G_0(A)\left(1 - G_0(A)\right)}{\alpha_0^2(\alpha_0 + 1)} = \frac{G_0(A)\left(1 - G_0(A)\right)}{\alpha_0 + 1}$$

۴- می دانیم اگر یک بردار تصادفی گاما داشته باشیم $z_i \sim \Gamma(\alpha_i, \theta)$ و بردار تصادفی π را از نرمال کردن این بردار به دست آوریم، $\left(\pi_i = \frac{z_i}{\sum_{i=1}^{k} z_i}\right)$ بردار تصادفی π از توزیع $(\pi_i)_{i=1\dots k} \sim Dir\left((\alpha_i)_{i=1\dots k}\right)$ پیروی می کند. از خواص توزیع گاما، داریم:

$$\left.\begin{array}{l} z_1 \sim \Gamma(\alpha_1, \theta) \\ z_2 \sim \Gamma(\alpha_2, \theta) \end{array}\right\} \xrightarrow[z_2]{z_1 \perp\!\!\!\perp z_2} (z_1 + z_2) \sim \Gamma(\alpha_1 + \alpha_2, \theta)$$

پس داریم:

$$\left(\pi_1 + \pi_2, \pi_3, \dots, \pi_k\right) = \frac{1}{\sum_{i=1}^{k} z_i} \times \left(z_1 + z_2, z_3, \dots, z_k\right)$$

$$= \frac{1}{\sum_{i=1}^{k} \Gamma(\alpha_i, 1)} \left(\Gamma(\alpha_1 + \alpha_2, 1), \Gamma(\alpha_3, 1), \dots, \Gamma(\alpha_k, 1)\right) \sim Dir\left(\alpha_1 + \alpha_2, \alpha_3, \dots, \alpha_k\right)$$

CE956: Statistical Learning
Department of Computer Engineering
Sharif University of Technology
Spring 2019: Room CE204, Sat. & Mon.: 13:30-15:00

Quiz 02 (20 Points) – (March-02-2019)

Solution

Gaussian Process: (10 points each)

1. Can object recognition be solved by a Gaussian Process (GP) Regression?
   - Yes. (Please refer to the class notes and reference papers).
     Given an image x, predict the class of the object present in the image y (i.e. y = {car, bus, bicycle). Although this is a classification task, one can treat the categories as real values and formulate the problem as regression and solve the regression problem with GP.

2. Given an overdetermined system of equations (y = mx+c) (i.e. 3 equations 2 unknowns), can we solve this system with GP?
   - Yes. (Please refer to the class notes and reference papers).
     This is also a regression problem. Just add some noise to the system of equations with the proper assumption and solve this regression problem with GP.

Stochastic Processes October 14, 2013
Sharif University of Technology CE 40-695
Dr. H.R. Rabiee

Time: 20 mins

Name: Std. Number:

Quiz 5 (Stochastic Processes)

Questions

1. **[6 Pts.]** Let $\{X(t), -\infty < t < \infty\}$ be a zero-mean, WSS, normal process with the autocorrelation function

$$R_x(\tau) = \begin{cases} 1 - \frac{|\tau|}{T} & -T \le \tau \le T, \\ 0 & o.w. \end{cases}$$

Let $\{X(t_i), i = 1, 2, ..., n\}$ be a sequence of n samples of the process taken at the time instants

$$t_i = i\frac{T}{2}, i = 1, 2, ..., n.$$

Find the mean and the variance of the sample mean

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X(t_i)$$

Solution:
Since $X(t)$ is zero-mean and stationary, we have
$E[X(t_i)] = 0$ and $R_x(t_i, t_k) = R_x(t_k - t_i) = R_x[(k-i)\frac{T}{2}]$
Thus

$$E[\hat{\mu}] = E[\frac{1}{n}\sum_{i=1}^{n} X(t_i)] = \frac{1}{n}\sum_{i=1}^{n} E[X(t_i)] = 0$$

and

$$Var(\hat{\mu}) = E[\hat{\mu}^2] - E^2[\hat{\mu}] = E[\hat{\mu}^2]$$
$$= E[\frac{1}{n}\sum_{i=1}^{n} X(t_i)][\frac{1}{n}\sum_{k=1}^{n} X(t_k)]$$
$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n} E[X(t_i)X(t_k)]$$
$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n} R_x[(k-i)\frac{T}{2}]$$

according to equation of$R_x(\tau)$

$$R_x((k-i)\frac{T}{2}) = \begin{cases} 1 & k = i, \\ \frac{1}{2} & |k-i| = 1, \\ 0 & |k-i| \ge 2 \end{cases}$$

Thus $Var(\hat{\mu}) = \frac{1}{n^2}[n * 1 + 2(n-1) * \frac{1}{2} + 0] = \frac{1}{n^2}[2n-1]$

2. **[4 Pts.]** Let $X(t)$ be a WSS zero-mean Gaussian process with autocorrelation function $R_x(\tau) = e^{-|\tau|}$. Further let A be a Gaussian random variable with mean 0 and variance 1 and independent of X. Determine whether or not the following process is mean ergodic. $Y(t) = X(t) + A$

Solution:

first of all, we chek if $Y(t)$ is WSS or not, and then we check $\lim_{T\to\infty} \frac{1}{T}\int_0^T C(\tau)d\tau \longrightarrow 0$

$E[X(t) + A] = E[X(t)] + E[A] = 0 + 0 = 0$

$R_{YY}(t_1, t_2) = E\{Y(t_1)Y(t_2)\} = E\{(X(t_1) + A).(X(t_2) + A)\} =$

$E\{X(t_1).X(t_2)\} + E\{A^2\} + E\{X(t_1).A\} + E\{X(t_2).A\} =$

$E\{X(t_1).X(t_2)\} + E\{A^2\} = R_{XX}(\tau) + 1 = e^{-|\tau|} + 1$

So $Y(t)$ is WSS.

$C_{YY}(t_1, t_2) = R_{YY}(t_1, t_2) - \mu_Y^2 = R_{YY}(\tau) = R_{XX}(\tau) + 1 = e^{-|\tau|} + 1$

$\Rightarrow C_{YY}(\tau) = e^{-|\tau|} + 1$

$\frac{1}{T}\int_0^T C_{YY}(\tau)d\tau = \frac{1}{T}\int_0^T (e^{-|\tau|} + 1)d\tau = \frac{1}{T}\int_0^T (e^{-\tau} + 1)d\tau = (\frac{-e^{-\tau}+\tau}{T})|_0^T = \frac{T-e^{-T}+1}{T}$

$\Rightarrow \lim_{T\to\infty} \frac{1}{T}\int_0^T C(\tau)d\tau = 1$

$\Rightarrow$ this process isn't mean ergodic!

Stochastic Processes October 27, 2014
Sharif University of Technology CE 40-695
H.R. Rabiee

Time: 20 mins

Name: Std. Number:

Quiz 5 (Power Spectrum + Ergodicity)

Questions

1. Consider the process
$$X(t) = ae^{j(wt-\phi)}$$
Where, a is a real constant, w is a random variable with density $f_w(w)$, and ϕ is a random variable independent of w and uniform in the interval $(0, 2\pi)$. We know that this process is WSS with zero mean. Find $S_x(w)$.

2. Let $X(t)$ be a WSS zero mean Gaussian process with autocorrelation $R_X(\tau) = 2e^{-2|\tau|}$. We define the process $Y(t) = Z(t) + X(t)$, Where $Z(t) = A$, and A is a Random Variable with mean μ_A and variance σ_A^2 and uncorrelated of $X(t)$. Speak about mean-ergodicity of $Y(t)$ in terms of μ_A and σ_A^2.

Solutions

1. We first define the autocorrelation function,
$R_X(\tau) = E[X(t+\tau)X^*(t)] = E[ae^{j(w(t+\tau)-\phi)}ae^{-j(wt-\phi)}] = a^2E[e^{jw\tau}]$
using the definition of expectation, and $f_w(w)$,
$R_X(\tau) = a^2\int_{-\infty}^{\infty} f_w(w)e^{jw\tau}dw$
From this and the uniqueness property of Fourier transform, it follows that:
$R_X(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_x(w)e^{jw\tau}dw$
Therefore,
$S_x(w) = 2\pi a^2 f_w(w)$

2. First we check if $Y(t)$ is WSS or not:
$E[Y(t)] = E[X(t)] + E[Z(t)] = 0 + \mu_A = \mu_A$
$R_Y(t_1, t_2) = E[Y(t_1)Y(t_2)] = E[(X(t_1) + Z(t_1))(X(t_2) + Z(t_2))]$
$= E[X(t_1)X(t_2)] + \mu_A E[X(t_1)] + \mu_A E[X(t_2)] + E[A^2]$
$= R_X(\tau) + E[A^2]$
So, $Y(t)$ is also WSS.
Now we check $\lim_{T\to\infty}\frac{1}{T}\int_0^T C_Y(\tau)d\tau \to 0$
$C_Y(\tau) = R_Y(\tau) - E^2[Y(t)] = R_X(\tau) + E[A^2] - \mu_A^2 = R_X(\tau) + \sigma_A^2$
$C_Y(\tau) = R_X(\tau) = 2e^{-2|\tau|} + \sigma_A^2$
$\lim_{T\to\infty}\frac{1}{T}\int_0^T C_Y(\tau)d\tau = \lim_{T\to\infty}\frac{1}{T}\int_0^T(2e^{-2|\tau|} + \sigma_A^2)d\tau$
$= \lim_{T\to\infty}\frac{-e^{-2\tau}+\tau\sigma_A^2|_0^T}{T} = \lim_{T\to\infty}\frac{1-e^{-2T}+T\sigma_A^2}{T} = \sigma_A^2$
The condition for $Y(t)$ to be mean ergodic is that $\sigma_A^2 = 0$. Since A is a R.V. With mean μ_A and variance σ_A^2, and the variance is equal to 0. This indicates that A should be a constant, and any R.V. with a variance $\sigma_A^2 \neq 0$ will result $Y(t)$ not to be mean ergodic.

Stochastic Processes October 1, 2012
Sharif University of Technology CE 40-695
Dr. H.R. Rabiee

Time: 20 mins

Name: Std. Number:

Quiz 3 (Stochastic Processes)

Questions

1. Suppose that $\mathbf{U} \sim uniform(0,1)$ and $\mathbf{X}(t) = t\mathbf{U} + 1$. [6 Pts.]

   (a) Find the first order p.d.f. of $\mathbf{X}(t)$.

   (b) Find $E[\mathbf{X}(\mathbf{t})]$ and $C_{XX}(t_1, t_2)$.

2. Let $\mathbf{X}(t)$ be a real-valued WSS gaussian random process with $E[\mathbf{X}(t)] = 1$ and $R_{XX}(\tau) = \frac{sin(\tau)}{\tau} + 1$. Show that for any fix t_c, $\mathbf{X}(t_c)$ and $\mathbf{X}(t_c + \pi)$ are independent. [4 Pts.]
   (*Hint:* A process $\mathbf{X}(t)$ is called gaussian, if the random variables $\mathbf{X}(t_1), ..., \mathbf{X}(t_n)$ are jointly gaussian for any n and $t_1, ..., t_n$.)

Stochastic Processes October 1, 2012
Sharif University of Technology CE 40-695
Dr. H.R. Rabiee

Time: 20 mins

Name: Std. Number:

Quiz 3 (Stochastic Processes)

Questions

1. Suppose that $\mathbf{U} \sim uniform(0,1)$ and $\mathbf{X}(t) = t\mathbf{U} + 1$. [6 Pts.]

   (a) Find the first order p.d.f. of $\mathbf{X}(t)$.

   (b) Find $E[\mathbf{X}(\mathbf{t})]$ and $C_{XX}(t_1, t_2)$.

   Solution:
   a)

$$f_X(x,t) = \frac{f_U(u)}{|g'(u)|}$$
$$where \quad u = \frac{x-1}{t}$$

   So:

$$f_{\mathbf{X}}(x;t) = \frac{1}{|t|} f_{\mathbf{U}}(\frac{x-1}{t}) = \begin{cases} \frac{1}{|t|} & 1 \le x \le t+1 \\ 0 & o.w. \end{cases}$$

   b)

$$E[\mathbf{X}(t)] = E[t\mathbf{U}+1]$$
$$= tE[\mathbf{U}] + 1 = \frac{t}{2} + 1$$

$$R(t_1,t_2) = E[\mathbf{X}(t_1)\mathbf{X}(t_2)] = E[(t_1\mathbf{U}+1)(t_2\mathbf{U}+1)]$$
$$= t_1t_2E[\mathbf{U}^2] + (t_1+t_2)E[\mathbf{U}] + 1 = \frac{t_1t_2}{3} + \frac{t_1+t_2}{2} + 1$$

$$C_{XX}(t_1,t_2) = R(t_1,t_2) - E[\mathbf{X}(t_1)]E[\mathbf{X}(t_2)]$$
$$= \frac{t_1t_2}{3} + \frac{t_1+t_2}{2} + 1 - [(\frac{t_1}{2}+1)(\frac{t_2}{2}+1)] = \frac{t_1t_2}{12}$$

2. Let $\mathbf{X}(t)$ be a real-valued WSS gaussian random process with $E[\mathbf{X}(t)] = 1$ and $R_{XX}(\tau) = \frac{sin(\tau)}{\tau} + 1$. Show that for any fix t_c, $\mathbf{X}(t_c)$ and $\mathbf{X}(t_c + \pi)$ are independent. [4 Pts.]
   (*Hint:* A process $\mathbf{X}(t)$ is called gaussian, if the random variables $\mathbf{X}(t_1), ..., \mathbf{X}(t_n)$ are jointly gaussian for any n and $t_1, ..., t_n$.)

Solution:

$$\begin{aligned} C(t_c, t_c + \pi) &= E[\mathbf{X}(t_c)\mathbf{X}(t_c + \pi)] - E[\mathbf{X}(t_c)]E[\mathbf{X}(t_c + \pi)] \\ &= R_{XX}(\pi) - 1 = 0 \end{aligned}$$

So $\mathbf{X}(t_c)$ and $\mathbf{X}(t_c + \pi)$ are uncorrelated. Since $\mathbf{X}(t)$ is a Gaussian random process, $\mathbf{X}(t_c)$ and $\mathbf{X}(t_c + \pi)$ are jointly gaussian and thus they are also independent.

Stochastic Processes October 19, 2015
Sharif University of Technology CE 40-695
H.R. Rabiee

Time: 20 mins

Name: Std. Number:

Quiz 5 (Poisson and Gaussian Processes)

Questions

1. If events are occurring randomly in time, and $N(t)$ denotes the number of events that occur by time t, then we say that $\{N(t), t \geq 0\}$ constitutes a *nonhomogeneous Poisson process* with intensity function $\lambda(t)$, $t \geq 0$, if

   - $N(0) = 0$
   - The numbers of events that occur in disjoint time intervals are independent.
   - $\lim_{h \to 0} \mathbb{P}[\text{ exactly 1 event between } t \text{ and } t+h]/h = \lambda(t)$.
   - $\lim_{h \to 0} \mathbb{P}[\text{ 2 or more events between } t \text{ and } t+h]/h = 0$.

   If we define

   $$m(t) = \int_0^t \lambda(s)\, ds\ ,\ t \geq 0$$

   we can prove that $N(t+s) - N(t)$ is a poisson random variable with mean $m(t+s) - m(t)$. Let $T_1, T_2, \ldots$ denote the interarrival times of events of a nonhomogeneous Poisson process having intensity function $\lambda(t)$.

   - (3 pts) Are T_is independent?
   - (4 pts) Find distribution of T_1 and T_2.

   [Hint: Poisson process has pmt:$\frac{\lambda^k}{k!}e^{-\lambda}$, mean: λ, variance: λ]

   [Hint2: Interarrival is time between two consecutive events]

   Solution.

   - No. For example for T_1 and T_2 we have:

     $$\begin{aligned}\mathbb{P}(T_2|T_1 = s) &= \mathbb{P}(0 \text{ events in} (s, s+t]|T_1 = s) \\ &= \mathbb{P}(0 \text{ events in} (s, s+t]) \\ &= e^{-(m(t+s)-m(s))}\end{aligned}$$

     which depends on s. So T_2 depends on T_1.

- We can write

$$F_{T_1} = \mathbb{P}(T_1 \leq t) = 1 - \mathbb{P}(T_1 > t) = 1 - \mathbb{P}(N(t) = 0) = 1 - e^{-m(t)}$$

so the density function of T_1 is

$$f_{T_1} = \frac{d}{dt} F_{T_1}(t) = \lambda(t) e^{-m(t)}$$

The distribution of T_2 can be derived using f_{T_1}

$$\begin{aligned}
\mathbb{P}(T_2 > t) &= \int_0^\infty \mathbb{P}(T_2 > t | T_1 = s) f_{T_1} \, ds \\
&= \int_0^\infty e^{-(m(t+s)-m(s))} \lambda(s) e^{-m(s)} \, ds \\
&= \int_0^\infty \lambda(s) e^{-m(t+s)} \, ds
\end{aligned}$$

Thus the density of T_2 is

$$f_{T_2}(t) = -\frac{d}{dt} \mathbb{P}(T_2 > t) = \int_0^\infty \lambda(s) \lambda(t+s) e^{-m(t+s)} \, ds$$

2. (3 pts) Consider following gaussian process:

$$X_0 \sim N(0, \sigma^2)$$
$$X_n = \frac{1}{2} X_{n-1} + Z_n \ , \ n \geq 1$$

Where $Z_1, Z_2, Z_3, \ldots$ are i.i.d. $N(0,1)$ and independent of X_0. Find σ such that X_n is SSS.[Hint: You can assume that there is some value for σ that this process is SSS]

Solution. This process is gaussian so it is SSS iff it is WSS. Simply we see that independent of σ:

$$\mathbb{E}[X_n] = 0$$

For X_n to be stationary, necessarily it's variance must be independent of n so we can write:

$$\mathbb{E}[X_n^2] = \frac{1}{4}\mathbb{E}[x_{n-1}^2 + \mathbb{E}[Z_n^2] + \mathbb{E}[X_{n-1}Z_n] = \frac{1}{4}\mathbb{E}[X_n^2] + 1$$

Therefore $\sigma^2 = \mathbb{E}[x_n^2] = \frac{4}{3}$. Using this value for σ^2 we see that:

$$R_X[m,n] = \frac{4}{3} 2^{-|m-n|}$$

Stochastic Processes Fall, 2019
Sharif University of Technology CE 40-695
H.R. Rabiee

Time: 20 mins

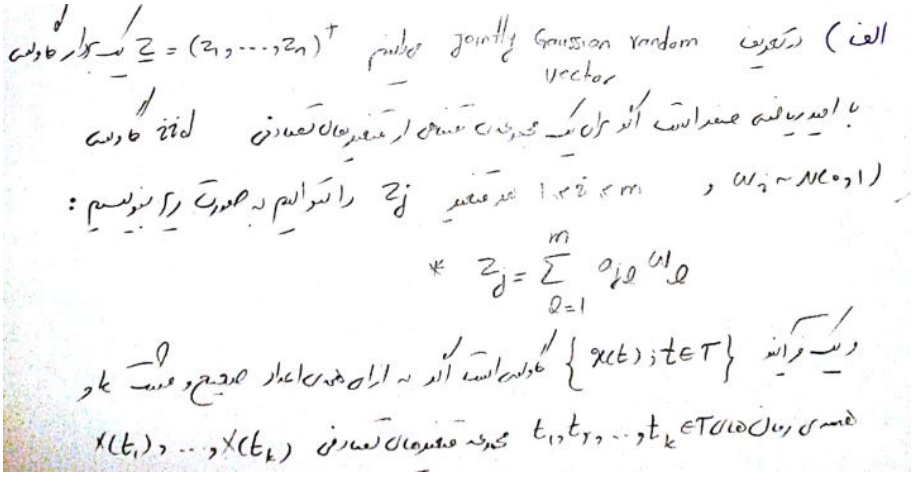
Name: Std. Number:

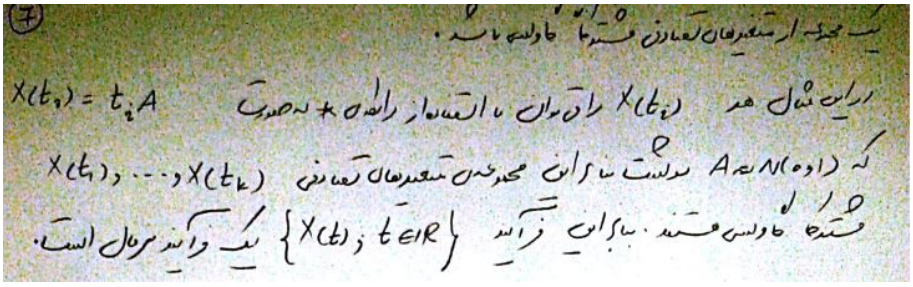
Quiz 6 (Gaussian Processes + Sufficient Statistic)

Questions

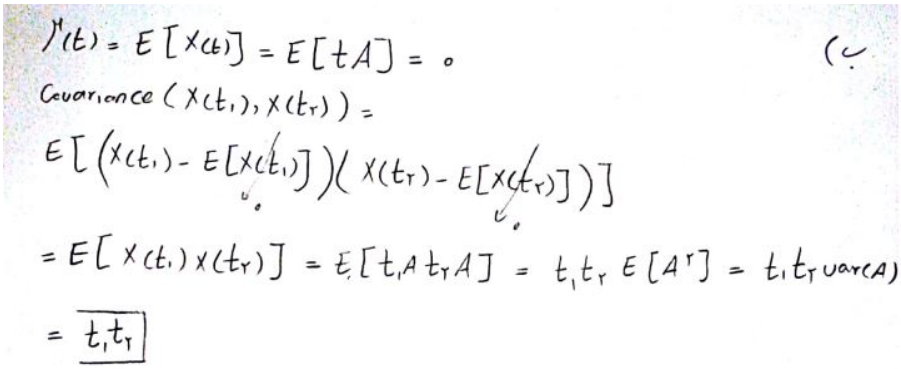
1. Consider Process $\{X(t); t \in R\}$ that $X(t) = At$ and $A \sim N(0, 1)$

   (a) Show that X(t) is a Gaussian process.





   (b) Find its expected value and covariance.



2. Suppose that $(X, Y) \sim^{iid} f(x, y|\theta) = \frac{2}{\theta^2}\mathbb{1}\{x, y \geq 0, x + y < \theta\}$. Find a sufficient statistic for θ.

   Solution: We can use factorization theorem. So we start with $f(X, Y|\theta)$:

$$f(X, Y|\theta) = \frac{2}{\theta^2}\mathbb{1}\{\forall_{0\leq i\leq n}\ x_i + y_i < \theta, x_i \geq 0, y_i \geq 0\}$$

$$= \frac{2}{\theta^2}\mathbb{1}\{\max_i\{x_i + y_i\} < \theta\}\mathbb{1}\{\forall_{0\leq i\leq n}\ x_i \geq 0, y_i \geq 0\}$$

   If we define $T(X, Y) = \max_i\{x_i + y_i\}$. So we can factorize $f(X, Y|\theta)$ into $h(X, Y)$ and $g(T(X, Y)|\theta)$ easily:

$$h(X, Y) = \mathbb{1}\{\forall_{0\leq i\leq n}\ x_i \geq 0, y_i \geq 0\}$$

$$g(T(X, Y)|\theta) = \frac{2}{\theta^2}\mathbb{1}\{T(X, Y) < \theta\}$$

   So according to factorization theorem, $T(X, Y) = \max_i\{x_i + y_i\}$ is a sufficient statistic for parameter θ.

به نام خدا

فرآیندهای تصادفی نیمسال اول ۹۹-۹۸

آزمونک ۵

سوال ۱:

اگر $X(t)$ یک فرآیند WSS نرمال با میانگین ۰ و تابع خودهمبستگی $R_X(\tau) = e^{-|\tau|}$ و A متغیر تصادفی نرمال با میانگین ۰ و کواریانس ۱ و مستقل از X باشد، آیا $Y(t) = X(t) + A$ فرآیند mean-ergodic است؟چرا؟

سوال ۲:

۱. فرآیند نقطه‌ای تکاملی [۱] چگونه فرآیندی است؟ توضیح دهید.

۲. تاریخچه H_t در این فرآیند شامل چه زمان‌هایی می‌شود؟

۳. می‌دانیم $f^*(t) = f(t|H_t)$ تابع چگالی شرطی برای زمان وقوع رویداد بعدی به شرط داشتن تاریخچه‌ی رویدادهای گذشته است. همچنین می‌دانیم تابع شدت شرطی [۲] با رابطه‌ی ۱ مشخص می‌شود.

$$\lambda^*(t) = \frac{f^*(t)}{۱ - F^*(t)} \quad (۱)$$

و همچنین داریم:

$$\lambda^*(t)dt = \mathbb{E}[N(dt)|H_t] \quad (۲)$$

که در رابطه‌ی اخیر dt بازه‌ای بینهایت کوچک در نزدیکی زمان t است. (با فرض این که در یک بازه‌ی زمانی بینهایت کوچک حداکثر یک رویداد به وقوع می‌پیوندد)

با توجه به این توضیحات، تابع شدت شرطی را تفسیر کنید.

سوال ۳:

در یکی از پاساژهای تهران یک مغازه‌ی لوازم التحریر فروشی وجود دارد و نرخ ورود مشتریان به این مغازه از یک فرآیند پواسون با نرخ دو مشتری در دقیقه پیروی می‌کند.

۱.۱. امید ریاضی مدت زمانی که طول می‌کشد تا بعد از باز شدن مغازه اولین مشتری وارد آن شود را به دست آورید.

۲.۱. با فرض این که در ۵ دقیقه‌ی آغازین شروع به کار مغازه، تنها یک مشتری وارد آن شده است، احتمال این که این مشتری در دقیقه‌ی اول وارد مغازه شده باشد را بیابید.

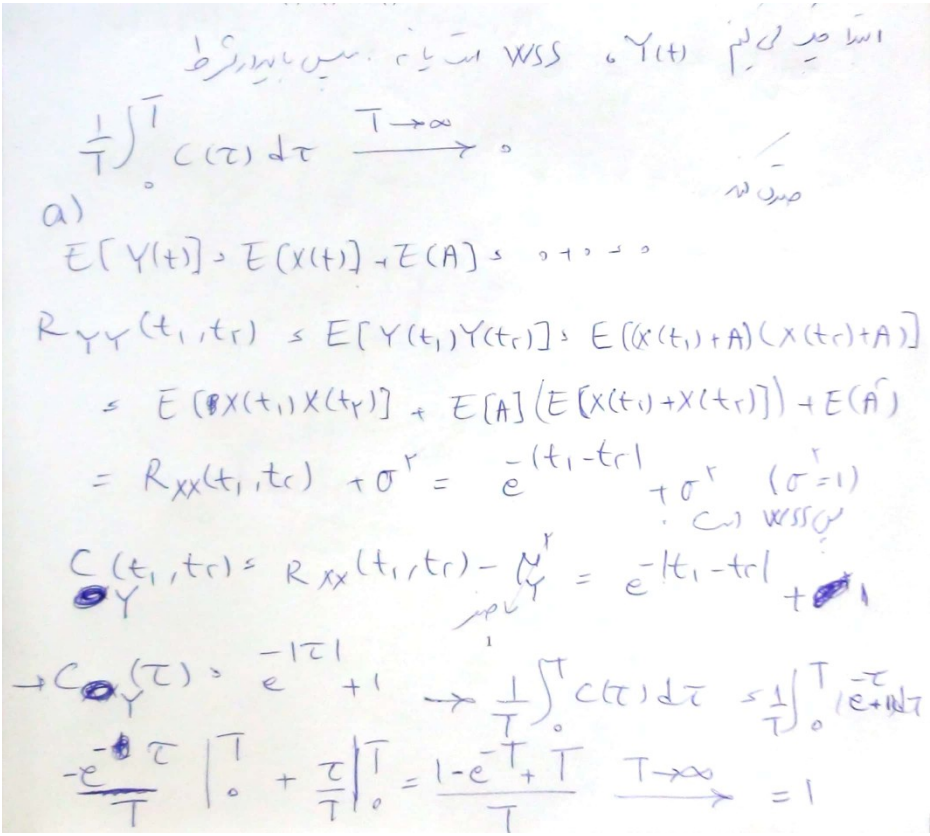
۳.۱. احتمال این که در ۵ دقیقه آغازین شروع به کار مغازه دقیقا سه مشتری وارد آن شده باشند را بیابید.

---

[۱] Evolutionary Point Process

[۲] Conditional Intensity Function

سوال ۱



سوال ۲

Evolutionarity

Usually we think of time as having an *evolutionary character*: what happens now may depend on what happened in the past, but not on what is going to happen in the future. This order of time is also a natural starting point for defining practically useful temporal point processes. Roughly speaking, we can define a point process by specifying a stochastic model for the time of the next event given we know all the times of previous events. The term *evolutionary point process* is used for processes defined in this way.

The past in a point process is captured by the concept of the *history* of the process. If we consider the time t, then the history $\mathcal{H}_t$ is the list of times of events $(\ldots, t_1, t_2, \ldots, t_n)$ up to but not including time t. Note that theoretically the point process may extend infinitely far back in time, but it does not have to do this. Note also that we assume that we have a *simple point process*, i.e. a point process where no points coincide, such that the points can be strictly ordered in time.

$$
\begin{aligned}
\lambda^*(t)\mathrm{d}t &= \frac{f^*(t)\mathrm{d}t}{1 - F^*(t)} \\
&= \frac{\mathbb{P}(\text{point in dt}|\mathcal{H}_t)}{\mathbb{P}(\text{point not before t}|\mathcal{H}_t)} \\
&= \frac{\mathbb{P}(\text{point in dt, point not before t}|\mathcal{H}_t)}{\mathbb{P}(\text{point not before t}|\mathcal{H}_t)} \\
&= \mathbb{P}(\text{point in dt}|\text{point not before t}, \mathcal{H}_t) \\
&= \mathbb{P}(\text{point in dt}|\mathcal{H}_t) \\
&= \mathbb{E}[N(\mathrm{d}t)|\mathcal{H}_t].
\end{aligned}
$$

Here $N(A)$ denotes the number of points falling in an interval, and the last equality follows from the assumption that no points coincide, so that there is either zero or one point in an infinitisemal interval. In other words, the conditional intensity function specifies the mean number of events in a region.

سوال ۳

3.(3x3) A basketball team scores baskets according to a Poisson process with rate 2 baskets per minute.

a) What is the expected amount of time until the team scores its first basket?

ANS: Let $N(t)$ be the number of baskets the team has made by time t and let T_1 be the time until the first basket. Since $N(t)$ is a Poisson process, its interarrival times, and T_1 in particular, are i.i.d. exponential RV's with parameter $\lambda = 2$. Thus $\mathbb{E}T_1 = 1/2$; i.e. expected time until first basket is half minute.

b) Given that at the five minute mark of the game the team has scored exactly one basket, what is the probability that the team scored the basket in the first minute?

ANS: Conditional on the event $\{N(t) = 1\}$, T_1 is uniformly distributed on $[0, t]$ (see Proposition 6.2.8). So $\mathbb{P}(T_1 < 1|N(5) = 1) = 1/5$.

c) What is the probability that the team scores exactly three baskets in the first five minutes of the game? (Computation of the numerical value is not necessary.)

ANS: For fixed t, $N(t)$ is a Poisson RV with parameter $\lambda t = 2t$; in particular $N(5)$ is Poisson with parameter 10. Thus

$$\mathbb{P}(N(5) = 3) = \frac{e^{-10}10^3}{3!}$$

Name: Std. Number:

Quiz 5 (Point Processes)

Questions

1. An arrival process is a sequence of increasing rv s, $0 < S_1 < S_2 < ...$, where $S_i < S_{i+1}$ means that $S_{i+1} - S_i$ is a positive rv, and S_i is the time when i^{th} event occurs. Any arrival process $\{S_n;\ n \geq 1\}$ can also be specified by either of two alternative stochastic processes. The first alternative is the sequence of interarrival times, $\{X_i; i \geq 1\}$. The X_i here are positive rv s defined in terms of the arrival epochs by $X_1 = S_1$ and $X_i = S_i - S_{i-1}$ for $i > 1$. The second alternative is the counting process $\{N(t); t > 0\}$, where for each $t > 0$ the rv N(t) is the aggregate number of arrivals up to and including time t.

   (a) Show that $P(S_n \leq t) = P(N(t) \geq n)$

   (b) Suppose $X_1, X_2, ...$ are iid rv s from $f_X(x) = \lambda e^{-\lambda x}$. We define a new rv $S_n = X_1 + X_2 + ... + X_n$ for all $n \geq 1$. Show that

$$f_{S_1,S_2,...,S_n}(s_1, s_2, ..., s_n) = \lambda^n e^{-\lambda s_n},\ for\ 0 < s_1 < s_2 < ... < s_n,\ and\ n > 1 \tag{1}$$

2. Given an unmarked point pattern $(t_1, ..., t_n)$ on an observation interval [0, T), show that the likelihood function is as follows where $\lambda^*(t)$ is the conditional intensity function at time t.

$$L = \left(\prod_{i=1}^{n} \lambda^*(t_i)\right) \exp\left(-\int_0^T \lambda^*(s)ds\right) \tag{2}$$

3. A final exam is started at time 0 for a class of n students. Each student is allowed to work until completing the exam. It is known that each students time to complete the exam is exponentially distributed with density $f_X(x) = \lambda e^{-\lambda x}; x \geq 0$. The times $X_1, ..., X_n$ are IID.

   (a) Let Z be the time at which the last student finishes. Show that Z has a distribution function $F_Z(z)$ given by $[1 - e^{-\lambda z}]^n$.

   (b) Let T_1 be the time at which the first student leaves. Show that the probability density of T_1 is given by $n\lambda e^{-n\lambda t}$. For each i, $2 \leq i \leq n$, let T_i be the interval from the departure of the $i-1^{\text{th}}$ student to that of the i^{th}. Show that the density of each T_i is exponential and find the parameter of that exponential density. Explain why T_is are independent.

Quiz-5 Solution (Point Processes)

Questions

1. (a) $\{Sn \leq t\}$ is the event that the n^{th} arrival occurs at some epoch $\tau \leq t$. This event implies that $N(\tau) = n$, and thus that $\{N(t) \geq n\}$. Similarly, $\{N(t) = m\}$ for some $m \geq n$ implies $\{S_m \leq t\}$, and thus that $\{S_n \leq t\}$.

   (b) for n=2:

$$f_{S_1,S_2}(s_1, s_2) = f_{X_1,S_2}(x_1, s_2) = f_{X_1}(x_1) f_{S_2|X_1}(s_2|x_1) = \lambda e^{-\lambda x_1} \times \lambda e^{-\lambda(s_2 - x_1)} = \lambda^2 e^{-\lambda s_2} \tag{1}$$

   Then we suppose we have:

$$f_{S_1,S_2,...,S_n}(s_1, s_2, ..., s_n) = \lambda^n e^{-\lambda s_n}, \ \textit{for some } n > 1 \tag{2}$$

   Then we can write:

$$\begin{aligned} f_{S_1,...,S_n,S_{n+1}}(s_1, ..., s_n, s_{n+1}) = \\ f_{S_1,...,S_n}(s_1, ..., s_n) f_{S_{n+1}|S_1,...,S_n}(s_{n+1}|s_1, ..., s_n) = \\ \lambda^n e^{-\lambda s_n} f_{S_{n+1}|S_1,...,S_n}(s_{n+1}|s_1, ..., s_n) \end{aligned} \tag{3}$$

   We know that $S_{n+1} = S_n + X_{n+1}$. So:

$$f_{S_{n+1}|S_1,...,S_n}(s_{n+1}|s_1, ..., s_n) = \lambda e^{-\lambda(s_{n+1} - s_n)} \tag{4}$$

   Then

$$\begin{aligned} f_{S_1,...,S_n,S_{n+1}}(s_1, ..., s_n, s_{n+1}) =& \lambda^n e^{-\lambda s_n} \lambda e^{-\lambda(s_{n+1} - s_n)} \\ =& \lambda^{n+1} e^{-\lambda s_{n+1}} \end{aligned} \tag{5}$$

2. The likelihood function is the joint density function of all the points in the observed point pattern $(t_1, ..., t_n) \in [0, T)$, and can therefore be factorised into all the conditional densities of each points given all points before it. This yields

$$L = f^*(t_1)...f^*(t_n)(1 - F^*(T)), \tag{6}$$

   where the last term $(1 - F^*(T))$ appears since the unobserved point $t_n + 1$ must appear after the end of the observation interval. So we can write:

$$\begin{aligned} L &= \left(\prod_{i=1}^{n} f^*(t_i) \right) \frac{f^*(T)}{\lambda^*(T)} \\ &= \left(\prod_{i=1}^{n} \lambda^*(t_i) \exp\left(- \int_{t_{i-1}}^{t_i} \lambda^*(s) ds \right) \right) \exp\left(- \int_{t_n}^{T} \lambda^*(s) ds \right) \\ &= \left(\prod_{i=1}^{n} \lambda^*(t_i) \right) \exp\left(- \int_{0}^{T} \lambda^*(s) ds \right) \end{aligned} \tag{7}$$

where t0 = 0.

3. (a) $Z \leq t$ if and only if $X_i \leq t$ for each i, $1 \leq n$, so

$$Pr\{Z \leq t\} = \prod_{i=1}^{n} Pr\{X_i \leq t\} = [1 - \exp(-\lambda t)]^n \tag{8}$$

(b) You can view T_1 as the time of the first arrival out of n Poisson processes each of rate λ. Thus T_1 is exponential with parameter $n\lambda$. More directly yet, $T_1 > t$ if and only if $X_i > t$ for $1 \leq i \leq t$, so $Pr\{T_1 > t\} = [exp(-\lambda t)]^n = exp(-n\lambda t)$. The time T_2 is the remaining time until the next student out of the remaining n-1 finishes. Because of the memorylessness of the exponential distribution, each of these n-1 students has an exponential time to go, so $Pr\{T_2 > t_2\} = \exp(-(n-1)\lambda t_2)$. Each of these times-to-go are independent of T_1, so T_2 is independent of T_1. In the same way T_i is exponential wih parameter $(n - i + 1)\lambda$ and is independent of the earlier T_is.

TH Quiz 5 (Pint Processes)

Due May 03, 2020 (11:59 pm)

1. Let N(t) be a Poisson point process with intensity λ=2, and let X1, X2, ⋯ be the corresponding inter-arrival times.
   a. Given that the third arrival occurred at time t=2, find the probability that the fourth arrival occurs after t=4.
   b. Consider the process at time t=10. Let T be the first arrival after t=10. Find E(T) and Var(T).
   c. If N(t) has rate λ, what is the distribution of arrival times T1, T2, ⋯. In particular, for n=1, 2, 3, ⋯, find $E[T_n]$ and $var(T_n)$.
   d. How do you generate the samples of arrival times by using i.i.d exponentially distributed random variables.

2. Assume we have two independent temporal point processes, with histories $H_1(t)$ and $H_2(t)$, using intensities $\lambda_1(t)$ and $\lambda_2(t)$, respectively. Let characterize the joint history $H(t) = H_1(t) \cup H_2(t)$, then find the corresponding intensity function λ(t) by setting up the proper differential equation.

Stochastic Processes October 19, 2015
Sharif University of Technology CE 40-695
H.R. Rabiee

Time: 20 mins

Name: Std. Number:

Quiz 5 (Poisson and Gaussian Processes)

Questions

1. If events are occurring randomly in time, and $N(t)$ denotes the number of events that occur by time t, then we say that $\{N(t), t \geq 0\}$ constitutes a *nonhomogeneous Poisson process* with intensity function $\lambda(t)$, $t \geq 0$, if

   - $N(0) = 0$
   - The numbers of events that occur in disjoint time intervals are independent.
   - $\lim_{h \to 0} \mathbb{P}[\text{ exactly 1 event between } t \text{ and } t+h]/h = \lambda(t)$.
   - $\lim_{h \to 0} \mathbb{P}[\text{ 2 or more events between } t \text{ and } t+h]/h = 0$.

   If we define

   $$m(t) = \int_0^t \lambda(s)\, ds\ ,\ t \geq 0$$

   we can prove that $N(t+s) - N(t)$ is a poisson random variable with mean $m(t+s) - m(t)$. Let $T_1, T_2, \ldots$ denote the interarrival times of events of a nonhomogeneous Poisson process having intensity function $\lambda(t)$.

   - (3 pts) Are T_is independent?
   - (4 pts) Find distribution of T_1 and T_2.

   [Hint: Poisson process has pmt:$\frac{\lambda^k}{k!}e^{-\lambda}$, mean: λ, variance: λ]

   [Hint2: Interarrival is time between two consecutive events]

   Solution.

   - No. For example for T_1 and T_2 we have:

     $$\begin{aligned}\mathbb{P}(T_2|T_1 = s) &= \mathbb{P}(0 \text{ events in} (s, s+t]|T_1 = s) \\ &= \mathbb{P}(0 \text{ events in} (s, s+t]) \\ &= e^{-(m(t+s)-m(s))}\end{aligned}$$

     which depends on s. So T_2 depends on T_1.

- We can write

$$F_{T_1} = \mathbb{P}(T_1 \leq t) = 1 - \mathbb{P}(T_1 > t) = 1 - \mathbb{P}(N(t) = 0) = 1 - e^{-m(t)}$$

so the density function of T_1 is

$$f_{T_1} = \frac{d}{dt} F_{T_1}(t) = \lambda(t) e^{-m(t)}$$

The distribution of T_2 can be derived using f_{T_1}

$$\begin{aligned}
\mathbb{P}(T_2 > t) &= \int_0^\infty \mathbb{P}(T_2 > t | T_1 = s) f_{T_1} \, ds \\
&= \int_0^\infty e^{-(m(t+s)-m(s))} \lambda(s) e^{-m(s)} \, ds \\
&= \int_0^\infty \lambda(s) e^{-m(t+s)} \, ds
\end{aligned}$$

Thus the density of T_2 is

$$f_{T_2}(t) = -\frac{d}{dt} \mathbb{P}(T_2 > t) = \int_0^\infty \lambda(s) \lambda(t+s) e^{-m(t+s)} \, ds$$

2. (3 pts) Consider following gaussian process:

$$X_0 \sim N(0, \sigma^2)$$
$$X_n = \frac{1}{2} X_{n-1} + Z_n \ , \ n \geq 1$$

Where $Z_1, Z_2, Z_3, \ldots$ are i.i.d. $N(0,1)$ and independent of X_0. Find σ such that X_n is SSS.[Hint: You can assume that there is some value for σ that this process is SSS]

Solution. This process is gaussian so it is SSS iff it is WSS. Simply we see that independent of σ:

$$\mathbb{E}[X_n] = 0$$

For X_n to be stationary, necessarily it's variance must be independent of n so we can write:

$$\mathbb{E}[X_n^2] = \frac{1}{4}\mathbb{E}[x_{n-1}^2 + \mathbb{E}[Z_n^2] + \mathbb{E}[X_{n-1}Z_n] = \frac{1}{4}\mathbb{E}[X_n^2] + 1$$

Therefore $\sigma^2 = \mathbb{E}[x_n^2] = \frac{4}{3}$. Using this value for σ^2 we see that:

$$R_X[m,n] = \frac{4}{3} 2^{-|m-n|}$$

Statistical Machine Learning
Sharif University of Technology
Hamid R. Rabiee

May 3, 2020
CE 40-956

TH-Quiz 6 (Point Processes)

Due May 12, 2020 (11:59 pm)

1. Suppose that $\{N_1(t), t \geq 0\}$ and $\{N_2(t), t \geq 0\}$ are independent Poisson processes with rates λ_1 and λ_2. Show that $\{N_1(t) + N_2(t), t \geq 0\}$ is a Poisson process with rate $\lambda_1 + \lambda_2$. Also, show that the probability that the first event of the combined process comes from $\{N_1(t), t \geq 0\}$ is $\lambda_1/(\lambda_1 + \lambda_2)$, independently of the time of the event.

2. Buses arrive at a certain stop according to a Poisson process with rateλ. If you take the bus from that stop then it takes a time R, measured from the time at which you enter the bus, to arrive home. If you walk from the bus stop the it takes a time W to arrive home. Suppose that your policy when arriving at the bus stop is to wait up to a time s, and if a bus has not yet arrived by that time then you walk home.

   (a) Compute the expected time from when you arrive at the bus stop until you reach home.

   (b) Show that if $W < 1/\lambda + R$ then the expected time of part (a) is minimized by letting $s = 0$; if $W > 1/\lambda + R$ then it is minimizes by letting $s = \infty$(that is, you continue to wait for the bus); and when $W = 1/\lambda + R$ all values of s give the same expected time.

   (c) Give an intuitive explanation of why we need only consider the cases s $= 0$ and $s = \infty$ when minimizing the expected time.

Statistical Machine Learning
Sharif University of Technology
Hamid R. Rabiee

CE 40-956

Solutions, TH-Quiz 6 (Point Processes)

Q1

Suppose that $\{N_1(t), t \geq 0\}$ and $\{N_2(t), t \geq 0\}$ are independent Poisson processes with rates λ_1 and λ_2. Show that $\{N_1(t)+N_2(t), t \geq 0\}$ is a Poisson process with rate $\lambda_1+\lambda_2$. Also, show that the probability that the first event of the combined process comes from $\{N_1(t), t \geq 0\}$ is $\lambda_1/(\lambda_1+\lambda_2)$, independently of the time of the event.

Solution: We check that $N(t) = N_1(t) + N_2(t)$ satisfies Definition 1.
(i) $N(t) = 0$.
(ii) Note that $N_1(t)$ and N_2 have independent increments. Moreover, $N_1(t)$ and $N_2(t)$ are independent.
(iii) Indeed, for any $t, s > 0$,

$$\begin{aligned}
\mathbb{P}(N(t+s) - N(t) = n) &= \sum_{k=0}^{n} \mathbb{P}(N_1(t+s) - N_1(t) = n-k | N_2(t+s) - N_2(t) = k) \\
&\quad \times \mathbb{P}(N_2(t+s) - N_2(t) = k) \\
&= \sum_{k=0}^{n} \frac{(\lambda_1 s)^{n-k}}{(n-k)!} \exp\{-\lambda_1 s\} \frac{(\lambda_2 s)^k}{k!} \exp\{-\lambda_2 s\} \\
&= \exp\{-(\lambda_1+\lambda_2)s\} \sum_{k=0}^{n} \frac{(\lambda_1 s)^{n-k}(\lambda_2 s)^k}{(n-k)!k!} \\
&= \frac{((\lambda_1+\lambda_2)s)^n}{n!} \exp\{-(\lambda_1+\lambda_2)t\}\,.
\end{aligned}$$

Now to show that the probability of the first arrival is from $N_1(t)$. Let X be the first arrival time for $N(t)$, and X_1, X_2 the corresponding times for $N_1(t)$ and $N_2(t)$.

One way to do is, observing that, $X \sim$ Exponential$(\lambda_1+\lambda_2)$,

$$\begin{aligned}
\mathbb{P}(\text{first event from } N_1(t) | X = x) &= \lim_{\delta_x \to 0} \mathbb{P}(X_1 < X_2 | X \in [x, x+\delta_x]) \\
&= \lim_{\delta_x \to 0} \frac{\mathbb{P}(X_1 \in [x, x+\delta_x])\mathbb{P}(X_2 > x) + o(\delta_x)}{\mathbb{P}(X \in [x, x+\delta_x])} \\
&= \frac{e^{-\lambda_1 x}(\lambda_1 \delta_x + o(\delta_x))e^{-\lambda_2 x} + o(\delta_x)}{e^{-(\lambda_1+\lambda_2)x}((\lambda_1+\lambda_2)\delta_x + o(\delta_x))} = \frac{\lambda_1}{\lambda_1+\lambda_2}\,.
\end{aligned}$$

As required, this probability does not depend on the first event time for $N(t)$.

Q2

Buses arrive at a certain stop according to a Poisson process with rate λ. If you take the bus from that stop then it takes a time R, measured from the time at which you enter the bus, to arrive home. If you walk from the bus stop the it takes a time W to arrive home. Suppose that your policy when arriving at the bus stop is to wait up to a time s, and if a bus has not yet arrived by that time then you walk home.

(a) Compute the expected time from when you arrive at the bus stop until you reach home.

(b) Show that if $W < 1/\lambda + R$ then the expected time of part (a) is minimized by letting $s = 0$; if $W > 1/\lambda + R$ then it is minimizes by letting $s = \infty$ (that is, you continue to wait for the bus); and when $W = 1/\lambda + R$ all values of s give the same expected time.

(c) Give an intuitive explanation of why we need only consider the cases $s = 0$ and $s = \infty$ when minimizing the expected time.

Solution: **(a)** Let $E_s = \mathbb{E}$(journey time for strategy s). The journey time is the function of the first arrival time of the rate λ Poisson process of bus arrivals. This has Exponential(λ) distribution (prop 2.2.1). So

$$E_s = \int_0^\infty \lambda e^{-\lambda t}[(t+R)\mathbf{1}(t \le s) + (s+W)\mathbf{1}(t > s)]\, dt$$

where $\mathbf{1}$ is the indicator function. Thus

$$\begin{aligned} E_s &= \int_0^s \lambda t e^{-\lambda t}\, dt + R\int_0^s \lambda e^{-\lambda t}\, dt + (s+W)\int_s^\infty \lambda e^{-\lambda t}\, dt \\ &= \frac{1-e^{-\lambda s}}{\lambda} + R(1-e^{-\lambda s}) + We^{-\lambda s} \end{aligned}$$

(b) Writing $E_s = (W - R - \frac{1}{\lambda})e^{-\lambda s} + \frac{1}{\lambda} + R$. We see that E_s is a decreasing function of s for $(W - R - 1/\lambda) > 0$, and increasing function for $(W - R - 1/\lambda) < 0$ and constant if $(W - R - 1/\lambda) = 0$.

(c) From the memoryless property of the exponential distribution , if it was worth waiting some time $s_0 > 0$ for a bus, and the bus has not arrived at s_0, then resetting time suggests that it must be worth waiting another s_0 time units. Thus, if the optimal s is positive, it must be infinite.

CE956: Statistical Learning
Department of Computer Engineering
Sharif University of Technology
Spring 2019: Room CE204, Sat. & Mon.: 13:30-15:00

Quiz 04 (20 Points) – (April-20-2019)

Solution

Point Processes: (each question 5 points)

1. Briefly, but concisely define a temporal point process? What is a Mark in a temporal point process? In general, is the assumption of mark being independent of the dynamic of the system, valid?
   - In statistics and probability theory, a point process or point field is a collection of mathematical points randomly located on some underlying mathematical space such as the real line. A temporal point process is a random process whose realizations consist of the times of isolated events. Marks are simply the features that are associated with the events. In general, the assumption of mark being independent of the dynamic of system it is not a valid assumption.

2. In general, the intensity functions are related to density/distribution functions:
   a. Then, why do we use intensity functions as building blocks of a Temporal Point Process?
   b. Name 4 popular intensity functions and by example, discuss which ones are suitable for modeling events in a social network.
   - (a) The densities need to integrate to 1 and it is difficult to combine timelines with density function. On the other hand, intensities only need to be nonnegative and it is easy to combine timeline with them. Therefore, they are more suitable to model and interpret the real system. (b) Homogeneous Poisson process, inhomogeneous Poisson process, terminating processes, and self-exciting processes. The self-exciting processes (such as Hawkes processes) are more suitable to model the history dependent events in social networks.

3. Briefly explain, how do you use Temporal Point Processes along with non-parametric Bayesian models for modeling content diffusion over social media?
   - Since temporal point processes are suitable for predicting the time of the next events and their marks, non-parametric Bayesian models can be utilized to solve them as a regression problem with proper priors.

4. Briefly explain how NetCode detect communities by using individual activities?
   - Since individual activities within a community has more effect on others in the same community we may use the effect of individuals on other community members in the same community to detect communities. In NetCode, the inter-community effect and structure are being considered, simultaneously (please refer to the reference paper and class notes/discussions)

Statistical Machine Learning
Sharif University of Technology
Hamid R. Rabiee

April 12, 2020
CE 40-956

TH Quiz 4 (Indian Buffet Process)

Due April 19, 2020 (11:59 pm)

1. Assume $\pi_1 \sim Poisson(\alpha_1)$ and $\pi_2 \sim Poisson(\alpha_2)$. Show that:

$$(\pi_1 + \pi_2) \sim Poisson(\alpha_1 + \alpha_2).$$

2. An Indian Buffet Process with parameter $\alpha = 4$ is running.

   (a) $Z_{n,k}$ indicates the presence of k'th feature (dish) in n'th sample (customer). Find probability distribution for number of non-zero elements in the 1'st, 2'nd, and 20'th rows of Z.

   (b) Implement IBP with $\alpha = 4$, run it for 1000 times (terminate after 20 customers entered) and then calculate $\sum_i Z_{1,i}$, $\sum_i Z_{2,i}$ and $\sum_i Z_{20,i}$. Plot the histogram of observed draws for these random variables in a chart. Do the results approve your calculations?

3. An improved version of IBP is introduced as two parameter IBP. The generative process of two parameter IBP is as follows:

   - First customer orders π_1 dishes; where $\pi_1 \sim Poisson(\alpha)$.
   - For all (n $>$ 1), n'th customer do two things:
     - Taste each existing dish with probability $\frac{m_k}{\beta+n-1}$; where β is the second parameter and m_k is number of previous customers who have tasted k'th dish.
     - Order π_n new dishes; where $\pi_n \sim Poisson(\frac{\alpha\beta}{\beta+n-1})$

   (a) Find probability distribution for the number of non zero elements in k'th row of Z matrix.

   (b) Find probability distribution for total number of dishes after n customers visit this buffet.

   (c) Implement two parameter IBP. Your implementation must include visualization of Z matrix for the first 50 customers. Run your implementation for all combinations of $\alpha \in \{5, 10, 20\}$ and $\beta \in \{1, 2, 4\}$. Visualize the results (Z matrices).

   (d) Try to interpret effects of each parameter on the results.

۱- می‌دانیم: $MGF_X(t) = E[e^{tX}]$ ← تابع مولد گشتاور متغیر تصادفی

لذا:

$$MGF_{\pi_1}(t) = E[e^{t\pi_1}] = \sum_{n=0}^{\infty} Pr(\pi_1 = n)\, e^{tn} = \sum_{n=0}^{\infty} \frac{\alpha_1^n e^{-\alpha_1}}{n!} \times e^{tn}$$

$$= e^{-\alpha_1} \times \sum_{n=0}^{\infty} \frac{(\alpha_1 e^t)^n}{n!} = e^{-\alpha_1} \times e^{(\alpha_1 e^t)} = \boxed{e^{\alpha_1 (e^t - 1)}}$$

$MGF_{\pi_2}(t) =$ به همین ترتیب $= \boxed{e^{\alpha_2 (e^t - 1)}}$

چون π_1 و π_2 مستقل هستند، تابع مولد گشتاور برای جمع آن‌ها، از ضرب MGF های آن‌ها به دست می‌آید. لذا:

$$MGF_{\pi_1 + \pi_2}(t) = MGF_{\pi_1}(t)\, MGF_{\pi_2}(t) = e^{(\alpha_1 + \alpha_2)(e^t - 1)} = MGF_{\pi_3}(t) \quad S.T.\ \pi_3 \sim Poisson(\alpha_1 + \alpha_2)$$

$$\Rightarrow \pi_1 + \pi_2 = \pi_3 \sim \boxed{Poisson(\alpha_1 + \alpha_2)}$$

۲-(a)

$\sum_i Z_{1,i}$ = طبق تعریف $= \boxed{Poisson(\alpha)}$

$\sum_i Z_{2,i} = \sum_{j=1}^{\pi_1} Z_{2,j} + Poisson(\frac{\alpha}{2}) = Poisson(\frac{\alpha}{2}) + Poisson(\frac{\alpha}{2}) = \boxed{Poisson(\alpha)}$

($Z_{2,j}$: هر درام پاد قال؛ π_1: تعداد ظرف‌های مشتری اول)

$\sum_i Z_{\ell,i} = \sum_{j=1}^{\pi_{\ell-1}} Z_{\ell,j} + Poisson(\frac{\alpha}{\ell}) = Poisson(\frac{\ell-1}{\ell}\alpha) + Poisson(\frac{\alpha}{n}) = \boxed{Poisson(\alpha)}$

۲-(b)- در فایل کدها نمودار رسم شده است.

۳-(a)

$\sum_i Z_{1,i}$ = طبق تعریف $= \boxed{Poisson(\alpha)}$

به دلیل وجود exchangability، هر مشتری می‌تواند با مشتری اول جابجا شود؛ درحالی که تأثیری بر توزیع نداشته باشد.

لذا:

$$\sum_i Z_{j,i} = \boxed{Poisson(\alpha)}$$

۳-(b)

ظرف‌های سفارش داده شده (جدید) توسط مشتری k ام $\rightarrow N_k$

$$N_1 \sim Poisson(\alpha)$$

$$N_k \sim Poisson(\frac{\alpha\beta}{\beta + k - 1})$$

تعداد کل ظرف‌های فعال بعد از ورود k مشتری $\rightarrow \sum_{i=1}^{k} N_i \sim Poisson(\alpha) + \dots + Poisson(\frac{\alpha\beta}{\beta+k-1})$

$=$ طبق نتیجه‌ی سوال اول $= \boxed{Poisson\left(\alpha\beta \sum_{i=1}^{k} \frac{1}{\beta + i - 1}\right)}$

۳-(c) و (d)- در فایل کدها پاسخ داده شده است.

Name: Std. Number:

Quiz 3 (Dirichlet Processes)

Questions

1. if $P \sim \mathrm{DP}(\alpha)$, then, for any measurable sets A and B show that:
   (a) $E[P(A)] = \bar{\alpha}(A)$
   (b) $\mathrm{var}(P(A)) = \frac{\bar{\alpha}(A)\bar{\alpha}(A^c)}{1+|\alpha|}$
   (c) $\mathrm{cov}(P(A), P(B)) = \frac{\bar{\alpha}(A \bigcap B) - \bar{\alpha}(A)\bar{\alpha}(B)}{1+|\alpha|}$

2. Assume a Dirichlet process prior, $DP(\alpha)$, for distribitions G on X.Show that for any measurable disjoint subsets A_1 and A_2 of X, $\mathrm{corr}(G(A_1), G(A_2))$ is negative. Is the negative correlation for random probabilities induced by the DP prior a restriction? Discuss.

3. Sequence of variables $X_1, X_2, X_3, \ldots, X_n$ is exchangeable if the joint distribution is invariant to permutation. An infinite sequence is infinitely exchangeable if any subsequence is exchangeable.
   (a) Show that CRP is infinitely exchangeable
   (b) Discuss the relationship of infinitely exchangeable to i.i.d

پاسخ کوییز سوم

Dirichlet Process

$P \sim DP(\alpha) \longrightarrow \bar{\alpha}$ is base measure

سوال ۱-

(a) افراز $\{A, A^c\}$ را در فضا sigma measurable در نظر بگیرید.

$$\rightarrow P(A, A^c) \sim Dir(\alpha(A), \alpha(A^c)) = Beta(\alpha(A), \alpha(A^c))$$

می دانیم که امید ریاضی توزیع بتا برابر است با:

$$X \sim Beta(\alpha, \beta) \longrightarrow \mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$$

$$\Rightarrow \mathbb{E}[P(A)] = \frac{\alpha(A)}{\alpha(A)+\alpha(A^c)} = \frac{\lambda \bar{\alpha}(A)}{\lambda \underbrace{(\bar{\alpha}(A)+\bar{\alpha}(A^c))}_{1}} = \bar{\alpha}(A)$$

(b)

$$X \sim Beta(\alpha, \beta) \longrightarrow Var[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$\Rightarrow Var[P(A)] = \frac{\alpha(A)\alpha(A^c)}{(\alpha(A)+\alpha(A^c))^2(\alpha(A)+\alpha(A^c)+1)} = \frac{\lambda^2 \bar{\alpha}(A)\bar{\alpha}(A^c)}{\lambda^2(\lambda+1)} = \frac{\bar{\alpha}(A)\bar{\alpha}(A^c)}{1+|\alpha|}$$

(c)

$$\Big(P(A\cap B), P(A-B), P(B-A), P(A\cup B)^c\Big) \sim Dir\Big(\alpha(A\cap B), \alpha(A-B), \alpha(B-A), \alpha((A\cup B)^c)\Big)$$

$$Cov[P(A), P(B)] = Cov[P(A\cap B)+P(A-B), P(A\cap B)+P(B-A)]$$

$\rightarrow$ cov of dirichlet $\leftarrow$

$$= Var[P(A\cap B)] + Cov[P(A-B), P(A\cap B)] + Cov[P(B-A), P(A\cap B)] + Cov[P(B-A), P(A-B)]$$

$$= \frac{\bar{\alpha}(A\cap B)(1-\bar{\alpha}(A\cap B))}{1+|\alpha|} + \frac{-\bar{\alpha}(A-B)\bar{\alpha}(A\cap B)}{1+|\alpha|} + \frac{-\bar{\alpha}(B-A)\bar{\alpha}(A\cap B)}{1+|\alpha|} + \frac{-\bar{\alpha}(B-A)\bar{\alpha}(A-B)}{1+|\alpha|}$$

$$\Big(\bar{\alpha}(A\cap B) = \bar{\alpha}(A) - \bar{\alpha}(A-B) = \bar{\alpha}(B) - \bar{\alpha}(B-A)\Big)$$

$$= \frac{\bar{\alpha}(A\cap B) - \bar{\alpha}(A)\bar{\alpha}(B)}{1+|\alpha|}$$

سوال 2 -

$$\text{Corr}(P(A_1), P(A_2)) = \frac{\bar{\alpha}(A_1 \cap A_2) - \bar{\alpha}(A_1)\bar{\alpha}(A_2)}{\sqrt{\text{Var}(P(A_1))\text{Var}(P(A_2))}} \overset{\text{disjoint}}{=} \frac{-\bar{\alpha}(A_1)\bar{\alpha}(A_2)}{\sqrt{\quad}} < 0$$

در فرآیند دیریشله، همبستگی هر دو مجموعه همواره منفی است. یعنی رفتار در توزیع احتمالاتی تصادفی که مورد انتظار ما است این است که هر چی که به مجموعه‌های نزدیک یکدیگر نسبت داده می‌شود با هم افزایش و کاهش پیدا کنند. اما همبستگی منفی عکس این انتظار را نشان می‌دهد. در واقع می‌توان گفت در فرآیند دیریشله توپولوژی فضا برای نسبت دادن جرم‌ها در نظر گرفته نشده است.

سوال ۳ - برای یک جایگشت از ورود مشتری‌ها می‌توان نوشت

$$P(X_1, \dots, X_n) = \frac{\left[\prod_{i=1}^{k} (n_i - 1)!\right] \alpha^{k-1}}{(\alpha+1)(\alpha+2)\cdots(\alpha+n-1)}$$

واضح است که این رابطه احتمال برای تمام جایگشت‌ها یکسان خواهد بود. پس می‌توان گفت هر دنباله n تایی exchangable است. <= infinitely exchangeable است

- هر دنباله iid یک دنباله exchangeable است. اما عکس این قضیه لزوماً درست نیست

برای مثال دستگاه چینی با ۲ مشتری را در نظر بگیرید.

Statistical Learning
Sharif University of Technology
H.R. Rabiee

March 2, 2019
CE 40-956

Name: Std. Number:

Quiz 4 (Dirichlet Process)

Questions

1. Stochastic process can be seen as an indexed collection of random variables. It can be considered as a collection of random variables $\{X_t\}_{t\in T}$ where T is the index set and for each t, X_t is a function from one measure space $(\Omega, \mathcal{F})$ to another measure space $(\Omega', \mathcal{F}')$. In this setting, how can we define Dirichlet Process? Define index set and domain and target measure spaces. (hint: see [**?**] and read about Kolmogorov extension theorem)
   Solution: The index set of a DP is a field of an arbitrary measure space. For example consider $(\mathcal{X}, \mathcal{A}, \alpha)$ as a measure space. For each $A \in \mathcal{A}$ consider a random variable P_A from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{C})$ where $\mathcal{C}$ is the Borel σ-field on $\mathbb{R}$. In this way P can be considered a stochastic process with index set $\mathcal{A}$. We will call it a Dirichlet Process if for any measurable partition of $\mathcal{X}$ like $(A_1, \ldots, A_k)$, the joint distribution of $(P_{A_1}, \ldots, P_{A_k})$ has dirichlet distribution with parameter $(\alpha(A_1), \ldots, \alpha(A_k))$. So P can be considered as a random probability measure on $(\mathcal{X}, \mathcal{A})$. The existence of $(\Omega, \mathcal{F})$ is proved by Kolmogorov existence theorem.

References

[1] Ferguson, Thomas S. "A Bayesian analysis of some nonparametric problems." The annals of statistics (1973): 209-230.

CE956: Statistical Learning
Department of Computer Engineering
Sharif University of Technology
Spring 2019: Room CE204, Sat. & Mon.: 13:30-15:00

Quiz 03 (35 Points) – (March-16-2019)

Solution

Non-Parametric Models: (5 points each)

1. Briefly explain what is the Dirichlet Distribution and Dirichlet Process.
   - (Please refer to the class notes and reference papers). The Dirichlet distribution is a family of continuous multivariate probability distributions parameterized by a vector α of positive reals. It is a multivariate generalization of the beta distribution. The Dirichlet distribution is the conjugate prior of the categorical distribution and multinomial distribution. A Dirichlet process is a probability distribution whose range is itself a set of probability distributions.

2. Why Dirichlet Process is important in the context of Bayesian Nonparametric Models.
   - (Please refer to the class notes and reference papers). A Dirichlet process is often used in Bayesian inference to describe the prior knowledge about the distribution of random variables, referring to how likely it is that the random variables are distributed according to one or another particular distribution.

3. What is the difference between CRP and IBP?
   - (Please refer to the class notes and reference papers). The Chinese restaurant process is a discrete-time stochastic process, analogous to seating customers at tables in a Chinese restaurant with an infinite number of circular tables, each with infinite capacity. The Indian buffet process (IBP) is a stochastic process defining a probability distribution over sparse binary matrices with a finite number of rows and an infinite number of columns.

4. What is the relation between distributions over the binary matrices and CRP and IBP?
   - (Please refer to the class notes and reference papers). We may use binary matrix representation for clustering. In this representation, rows are data points, and columns are clusters. Since each data point is assigned to one and only one cluster, rows sum to one. In this context, the Chinese restaurant process (CRP) is the distribution on partitions of the data induced by a Dirichlet Process Mixture (DPM) where the number of columns is countably infinite. Thus, we can think of the CRP as a distribution on such binary matrices (hard membership). We may think of a more general distribution on binary matrices (soft membership) where rows are data points, and columns are latent features. We can think of infinite binary matrices where each data point can now have multiple features, so the rows can sum to more than one. In other words, we assume there are multiple overlapping clusters, and each data point can belong to several clusters simultaneously. IBP corresponds to the latter case.

5. Why we need to perform sampling for inference purposes in the context of Bayesian Nonparametric Models?
   - (Please refer to the class notes and reference papers). In practice, computation of posteriors is not tractable and we need approximate inference methods to compute them. We utilize sampling methods for these approximations.

6. Briefly explain what is Gibbs Distribution.
   - (Please refer to the class notes and reference papers). Gibbs distribution is a probability distribution or probability measure that gives the probability that a system will be in a certain state as a function of that state's energy and the temperature of the system. It is also a frequency distribution of particles in a system.

7. Briefly but concisely explain what MCMC is?
   - (Please refer to the class notes and reference papers). Markov chain Monte Carlo (MCMC) methods comprise a class of algorithms for sampling from a probability distribution. MCMC constructs a Markov chain that has the desired distribution as its equilibrium distribution, and obtains a sample of the desired distribution by observing the chain after a number of steps. The more steps there are, the more closely the distribution of the sample matches the actual desired distribution.

Name: Std. Number:

Quiz 7 (interpretable Learning)

Questions

1. We have three interpretation algorithms and we want to evaluate their performance on a trained model for a classification task. The model is trained on a labeled image dataset to classify images into two categories (Each image belongs to exactly one of these categories.). Each interpretation algorithm gives an importance score, for each category, to pixels of all images. So, for the ath image, we have two score maps S_a^1 and S_a^2 that show the importance of each pixel of this image for category 1 and 2 respectively. Design a machine based quantitative measure (which reports a number) for evaluating these three interpretation algorithms. You can use and run the trained classifier.

2. Existing approaches in interpretation can be categorized into 2 categories: (1) Perturbation and forward propagation based methods. (2) Backpropagation based methods

   (a) Explain the general idea behind each category and discuss about their advantages and disadvantages.

   (b) (The saturation problem) Why can't simple perturbation and backpropagation based algorithms interpret the following model successfully?

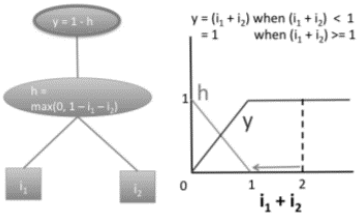


Figure 1:

   (c) To solve the above problem, instead of considering the derivation in backpropagation based models we can use the difference from a reference value. Imagine the target node t in a deep neural network. We define the $\Delta t = t - t_0$ in which t_0 is a reference value for example zero. We also define Δx_i for all other nodes x_i of the network. We define $C_{\Delta x_i, \Delta t}$ which shows the amount of Δt which is triggered by the Δx_i. So:

$$\Sigma_{i=1}^{N} C_{\Delta x_i, \Delta t} = \Delta t \tag{1}$$

We also define a multiplier $m_{\Delta x_i,\Delta t}$ as:

$$m_{\Delta x_i,\Delta t} = \frac{C_{\Delta x_i,\Delta t}}{\Delta x_i} \tag{2}$$

If y_js are a set of intermediate nodes between some nodes x_is and t, show that the chain rule (3) assumption is compatible with the equation (1). In fact given $C_{\Delta x_i,\Delta y_j}$ and $C_{\Delta y_j,\Delta t}$ both satisfy (1), you should show that defining $C_{\Delta x_i,\Delta t}$ according to the chain rule also satisfies (1).

$$m_{\Delta x_i,\Delta t} = \Sigma_j \; m_{\Delta x_i,\Delta y_j} \times m_{\Delta y_j,\Delta t} \tag{3}$$

(d) Explain how this new backpropagation technique can solve the saturation problem (part b).

Quiz 7 Solution (Interpretable Learning)

Solutions

1. Given an image that originally belongs to class 1, we identify which pixels to erase to convert the image to the target class 2. For each interpretation algorithm, we calculate a new score map $S_a = S_a^1 - S_a^2$ for each image and select top 20 percent pixels with highest scores and erase them. Then we obtain the classification accuracy of the predictive model on data with new features for class 1. The least accuracy belongs to the interpretation algorithm which gives highest scores to the most related pixels.

2. (a) Perturbation and forward propagation based methods change the input (omitting a feature, ...) and then compute the model output in a forward direction to asses the effect of the input variation on the model output. Its main disadvantage is that for each variation model should be run and it is very time consuming. Backpropagation based methods perform this task by computing the gradient of the output based on the input in a backward direction. It is limited to models in which the output is differentiable based on input. It also just considers a small neighborhood around the input. Read more about these techniques in section 2.1 and 2.2 of this paper `https://arxiv.org/pdf/1704.02685.pdf` .

   (b) For example in a situation in which $i_1 = 1$ and $i_2 = 1$ a backpropagation method calculates the gradient of the output based on the both input variables zero. Also in a forward based method a small change in variables do not change the output. So both of them report that the output is completely unrelated to both of variables which is obviously wrong.

   (c) Considering the target node z,

$$
\begin{aligned}
\sum_i C_{\Delta x_i \Delta z} &= \sum_i \Delta x_i m_{\Delta x_i \Delta z} \text{ (By definition of } m_{\Delta x_i \Delta z}) \\
&= \sum_i \Delta x_i \sum_j m_{\Delta x_i \Delta y_j} m_{\Delta y_j \Delta z} \text{ (By the chain rule)} \\
&= \sum_i \Delta x_i \sum_j \frac{C_{\Delta x_i \Delta y_j}}{\Delta x_i} m_{\Delta y_j \Delta z} \text{ (By definition of } m_{\Delta x_i \Delta y_j}) \\
&= \sum_i \sum_j C_{\Delta x_i \Delta y_j} m_{\Delta y_j \Delta z} \\
&= \sum_j \sum_i C_{\Delta x_i \Delta y_j} m_{\Delta y_j \Delta z} \text{ (Flipping the order of summation)} \\
&= \sum_j \Delta y_j m_{\Delta y_j \Delta z} \text{ (By summation-to-delta of } C_{\Delta x_i \Delta y_j}) \\
&= \sum_j \Delta y_j \frac{C_{\Delta y_j \Delta z}}{\Delta y_j} \text{ (By definition of } m_{\Delta y_j \Delta z}) \\
&= \sum_j C_{\Delta y_j \Delta z} = \Delta z \text{ (By summation-to-delta of } C_{\Delta y_j \Delta z})
\end{aligned}
$$

(d) In the case of (b), we can set the reference value of the target node y and both i_1 and i_2 zero. Now you can check and see that the problem is solved.

Name: Std. Number:

Quiz 8 (interpretable Learning)

Questions

1. We have a trained complex predictive model as a black box and we want to evaluate its performance on an N-dimensional input space. The input space is too large and it is divided into 10 homogeneous subsets. One labeled sample from each subset is available and the result of the running the model on one of these samples is unsatisfying.

   Interpret the behavior of the model in the locality of this sample by estimating the weight it gives to each feature. During your interpretation you can run the model or train and run any other predictive model.

Quiz 8 Solution (Interpretable Learning)

Solutions

1. We want to interpret the behavior of the model in the locality of that sample. We can select an arbitrary number of samples from this locality and run the predictive model on them to obtain the results. Now we have a set of (x, y) pairs by which we can train another model. We can select a linear model and train its weights by the train set that we constructed. Because these samples are obtained by the predictive model, these two models have a same behavior and weights of the linear model are a good estimation of the weights the black box model gives to features.

Statistical Machine Learning
Sharif University of Technology
Hamid R. Rabiee

May 27, 2020
CE 40-956

TH Quiz 8 (Interpretability)

Due June 11, 2020 (11:59 pm)

1. Below, we introduce two kind of interpretability technique title; first search about them. Indentify the model-specific or model-agnostic paramete of them and then use them to solve the problem of which should be used in following special cases in order to reduce complexity and making sutiatuion more interpretable.

   (a) Anchors

   (b) LOCO variable importance

   (c) LIME

   (d) Treeinterpreter

   (e) Shapley explanations

   Here's some situation. You should determine the usage of the disscused techniques for these cases (with explenations and ressons)

   i. Suppose a situation that we want to derive consistent local variable contributions to black-box model predictions. We know that our numbers are not large (or our trees are not deep) and there are some low-level codes.

   ii. Suppose a situation that we only have access to some most important local variables and we want to generate sparse or simplified, explanations using these variables. Totally we want to describe the average behavior of a complex machine-learned response function.

   iii. Suppose a situation that we want high precision and also we want to generates rules about the most important variables for a prediction.

2. A prediction can be explained by assuming that each feature value of the instance is a player in a game where the prediction is the payout. Shapley values tells us how to fairly distribute the payout among the features. Formally, the Shapley value is the average marginal contribution of a feature value across all possible coalitions.

   Suppose a linear model prediction for a single data point:

   $$\hat{f}(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

   Here x is the instance for which we want to compute the contributions. Each x_j is a feature value, with $j = 1, ..., p$. The $\beta_j$is the weight corresponding to feature j.
   The contribution is the difference between the feature effect minus the average effect. First, Calculate how much each feature contributed to the prediction.

   The Shapley value is defined via a value function val of players in S. The Shapley value of a feature value is its contribution to the payout, weighted and summed over all possible feature value combinations:

   $$\phi_j(val) = \sum_{S \subseteq \{x_1,\ldots,x_p\} \setminus \{x_j\}} \frac{|S|!(p-|S|-1)!}{p!} \left(val\left(S \cup \{x_j\}\right) - val(S)\right)$$

where S is a subset of the features used in the model, x is the vector of feature values of the instance to be explained and p the number of features. is the prediction for feature values in set S that are marginalized over features that are not included in set S:

$$val_x(S) = \int \hat{f}(x_1, \ldots, x_p) d\mathbb{P}_{x \notin S} - E_X(\hat{f}(X))$$

Now, suppose a concrete example: The machine learning model works with 4 features x_1, x_2, x_3 and x_4. Evaluate the prediction for the coalition S consisting of feature values x_1 and x_3 and compare it to feature contributions in the linear model. At the end, check if the Sharlpy value method satisfies the properties **Efficiency**, **Symmetry**, **Dummy** and **Additivity**. (Hint: Sharply value is a fair payout method)

Quiz 8 (Interpretability)

Solutions

1. Q1:

   Technique: Anchors
   Description: A newer approach from the inventors of LIME that generates high-precision sets of plain-language rules to describe a machine learning model prediction in terms of the model's input variable values.
   Suggested usage: Anchors is currently most applicable to classification problems in both traditional data mining and pattern-recognition domains. Anchors can be higher precision than LIME and generates rules about the most important variables for a prediction, so it can be a potential replacement for Shapley values for models that don't yet support the efficient calculation of Shapley values.
   Reference: Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "Anchors: High-Precision Model-Agnostic Explanations," The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), April 25, 2018,
   Global or local scope: Local.
   Best-suited complexity: Low to medium. Anchors can create explanations for very complex functions, but the rule set needed to describe the prediction can become large.
   Model specific or model agnostic: Model agnostic.

   Technique: LOCO variable importance
   Description: LOCO, or even LOFO, variously stands for leave-one-"column" or "covariate" or "feature"-out. LOCO creates local interpretations for each row in a training or unlabeled score set by scoring the row of data once and then again for each input variable (e.g., column, covariate, feature) in the row. In each additional scoring run, one input variable is set to missing, zero, its mean value, or another appropriate value for leaving it out of the prediction. The input variable with the largest absolute impact on the prediction for that row is taken to be the most important variable for that row's prediction. Variables can also be ranked by their impact on the prediction on a per-row basis.
   Suggested usage: You can use LOCO to build reason codes for each row of data on which nearly any complex model makes a prediction. LOCO can deteriorate in accuracy when complex nonlinear dependencies exist in a model. Shapley explanations might be a better technique in this case, but LOCO is model agnostic and has speed advantages over Shapley both in training and scoring new data.
   Reference: Jing Lei et al., "Distribution-Free Predictive Inference for Regression," arXiv:1604.04173, 2016,
   Global or local scope: Local but can be aggregated to create global explanations.
   Best-suited complexity: Any. LOCO measures are most useful for nonlinear, nonmonotonic

response functions but can be applied to many types of machine-learned response functions.
Model specific or model agnostic: Model agnostic.

Technique: LIME
Description: Typically uses local linear surrogate models to explain regions in a complex machinelearned response function around an observation of interest.
Suggested usage: Local linear model parameters can be used to describe the average behavior of a complex machine-learned response function around an observation of interest and to construct reason codes. LIME is approximate, but has the distinct advantage of being able to generate sparse, or simplified, explanations using only the most important local variables. Appropriate for pattern recognition applications as well. The original LIME implementation may sometimes be inappropriate for generating explanations in real-time on unseen data."
Reference: Ribeiro et al., "'Why Should I Trust You?' Explaining the Predictions of Any Classifier."
Best-suited complexity: Low to medium. Suited for response functions of high complexity but can fail in regions of extreme nonlinearity or high-degree interactions.
Global or local scope: Local.
Model specific or model agnostic: Model agnostic.

Technique: Treeinterpreter
Description: For each variable used in a model, treeinterpreter decomposes some decision tree, random forest, and GBM predictions into bias (overall training data average) and component terms. Treeinterpreter simply outputs a list of the bias and individual variable contributions globally and for each record.
Suggested usage: You can use treeinterpreter to interpret some complex tree-based models, and to create reason codes for each prediction. If you would like to use treeinterpreter, make sure your modeling library is fully supported by treeinterpreter. In some cases, treeinterpreter may not be locally accurate (local contributions do not sum to the model prediction) and treeinterpreter does not consider how contributions of many variables affect one another as carefully as the Shapley approach. However, treeinterpreter can generate explanations quickly. Also most treeinterpreter techniques appear as Python packages.
Reference: Ando Saabas, "Random Forest Interpretation with scikit-learn," Diving into Data [blog], August 12, 2015
Global or local scope: Local but can be aggregated to create global explanations.
Best-suited complexity: Any. Treeinterpreter is meant to explain the usually nonlinear, nonmonotonic response functions created by certain decision tree, random forest, and GBM algorithms.
Model specific or model agnostic: Treeinterpreter is model specific to algorithms based on decision trees

Technique: Shapley explanations
Description: Shapley explanations are a Nobel-laureate technique with credible theoretical support from economics and game theory. Shapley explanations unify approaches such as LIME, LOCO, and treeinterpreter to derive consistent local variable contributions to black-box model predictions. Shapley also creates consistent, accurate global variable importance measures.
Suggested usage: Shapley explanations are accurate, local contributions of input variables and can be rank-ordered to generate reason codes. Shapley explanations have long-standing theoretical support, which might make them more suitable for use in regulated industries, but they

can be time consuming to calculate, especially outside of decision trees in H2O.ai, LightGBM, and XGBoost where Shapley is supported in low-level code and uses the efficient Tree SHAP approach.
Reference: Lundberg and Lee, "A Unified Approach to Interpreting Model Predictions."
Global or local scope: Local but can be aggregated to create global explanations.
Best-suited complexity: Low to medium. This method applies to any machine learning model, including nonlinear and nonmonotonic models, but can be extremely slow for large numbers of variables or deep trees.
Model specific or model agnostic: Can be both. Uses a variant of LIME for modelagnostic explanations. Takes advantage of tree structures for decision tree models and is recommended for tree-based models

2. Q2

$$\phi_j(\hat{f}) = \beta_j x_j - E(\beta_j X_j) = \beta_j x_j - \beta_j E(X_j)$$

$$\begin{aligned}\sum_{j=1}^{p} \phi_j(\hat{f}) =& \sum_{j=1}^{p} (\beta_j x_j - E(\beta_j X_j)) \\ =& (\beta_0 + \sum_{j=1}^{p} \beta_j x_j) - (\beta_0 + \sum_{j=1}^{p} E(\beta_j X_j)) \\ =& \hat{f}(x) - E(\hat{f}(X))\end{aligned}$$

So, the evaluation for four features is:

$$val_x(S) = val_x(\{x_1, x_3\}) = \int_{\mathbb{R}} \int_{\mathbb{R}} \hat{f}(x_1, X_2, x_3, X_4) d\mathbb{P}_{X_2 X_4} - E_X(\hat{f}(X))$$

Efficiency:
The feature contributions must add up to the difference of prediction for x and the average.

$$\sum_{j=1}^{p} \phi_j = \hat{f}(x) - E_X(\hat{f}(X))$$

Symmetry:
The contributions of two feature values j and k should be the same if they contribute equally to all possible coalitions. If

$$val(S \cup \{x_j\}) = val(S \cup \{x_k\})$$

for all

$$S \subseteq \{x_1, \ldots, x_p\} \setminus \{x_j, x_k\}$$

then, $\phi_j = \phi_k$

Dummy
A feature j that does not change the predicted value – regardless of which coalition of feature values it is added to – should have a Shapley value of 0. If

$$val(S \cup \{x_j\}) = val(S)$$

for all

$$S \subseteq \{x_1, \ldots, x_p\}$$

then $\phi_j = 0$

Additivity:
For a game with combined payouts $val + val^+$ the respective Shapley values are as follows:

$$\phi_j + \phi_j^+$$

Suppose you trained a random forest, which means that the prediction is an average of many decision trees. The Additivity property guarantees that for a feature value, you can calculate the Shapley value for each tree individually, average them, and get the Shapley value for the feature value for the random forest.

CE956: Statistical Learning
Department of Computer Engineering
Sharif University of Technology
Spring 2019: Room CE204, Sat. & Mon.: 13:30-15:00

Quiz 06 (20 Points) – (May-18-2019)

Solution

Interpretable Learning:

1. What are the main deficiencies of deep networks? Explain. (5 points)
   - They are vulnerable to spoofing (adversarial attack problem)
   - They are inefficient (small data problem)
   - They lack to explain why? (interpretability problem)
   - They fail to explain common sense and functions
2. What is meta RL? In what respect it helps interpretability? (7.5 points)
   - Meta-learning algorithm try to automatically find the best architecture and hyperparameters in a learning problem (learning-to-learn: self-supervised learning). Meta RL is just applying meta-learning to RL by expanding the scope of training and training the network for more than one task by exploiting recurrent connections. Since it learns from the past and the learning process id self-supervised, it can reason and interpret to generate new concepts.
3. What is a graph network? In what respect it helps interpretability? (7.5 points)
   - The human prior knowledge are hardwired in brain (inductive biases). Graph neural networks are deep learning systems that have an innate bias toward representing things as objects and relations (the input to these systems are graphs of relations instead of signals). In other words, graph neural networks (GNNs) are connectionist models that capture the dependence of graphs via message passing between the nodes of graphs. Unlike standard neural networks, graph neural networks retain a state that can represent information and hence they are interpretable by predicting how relationships evolves over time.

Name: Std. Number:

Quiz 6 (Big Data Analytics)

Questions

1. Design MapReduce algorithm to get two matrices as input and output their multipication. Define exactly Map and Reduce functions.

   http://www.mathcs.emory.edu/ cheung/Courses/554/Syllabus/9-parallel/matrix-mult.html

2. Design MapReduce algorithm for BFS on graph and use it to find shortest path from a source node.

   http://www.cs.kent.edu/ jin/Cloud12Spring/GraphAlgorithms.pptx

3. Compare Hadoop and Spark. Describe advantages and disadvantages of each one.

   https://www.datamation.com/data-center/hadoop-vs.-spark-the-new-age-of-big-data.html

Statistical Learning
Sharif University of Technology
H.R. Rabiee

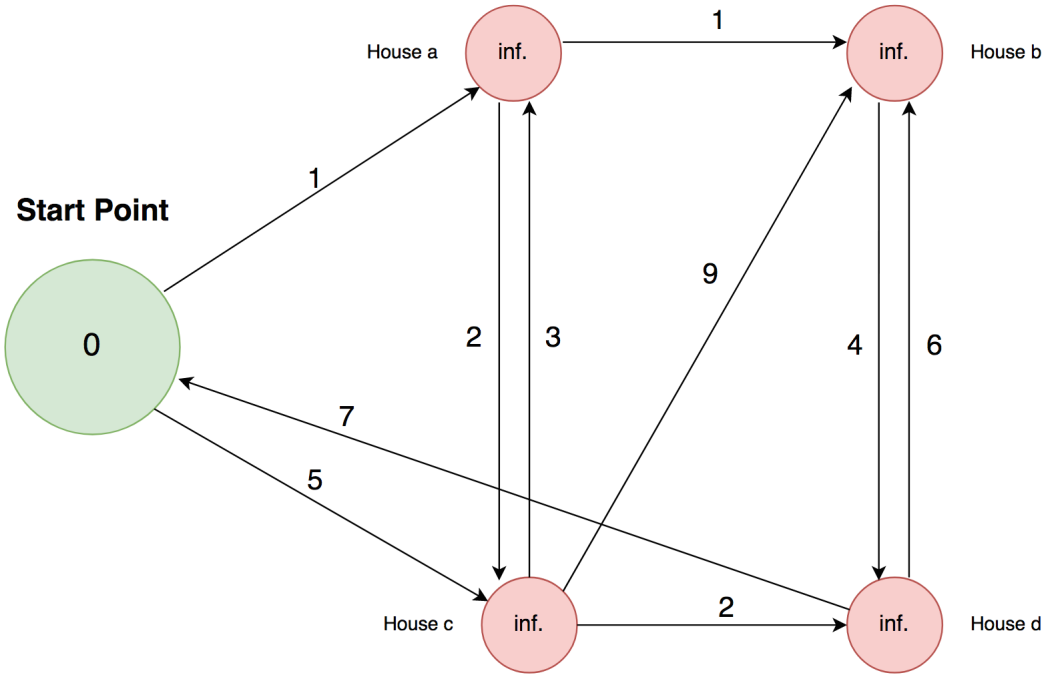
Due March 2, 2020
CE 40-956

Name: Std. Number:

Solutions, TH-Quiz 7 (Big data analysis)

Questions

1. A postman wants to move from the start point and deliver several letters to the houses, the location map is shown in the figure below. Use the Map Reduce algorithm to get the shortest path from the starting point to each house. First define exactly Map and Reduce functions and then solve this problem.



Input Data Format:

$$Node: < id, costFromSource, prevHopFromSource, AdjacencyList >$$
$$AdjacencyList: \{neighborNode, costToNeighborNode\}$$

Initial Input Data:

```
Node s: <s, 0, - , {(Node a, 1), (Node c, 5)}>
Node a: <a, ∞, - , {(Node b, 1), (Node c, 2)}>
Node b: <b, ∞, - , {(Node d, 4)}>
```

```
Node c: <c, ∞, - , {(Node a, 3), (Node b, 9), (Node d, 2)}>

Node d: <d, ∞, - , {(Node s, 7), (Node b, 6)}>
```

MapReduce Iteration Mapping:

$$map1: Node.id: Node \rightarrow \{list(Node.neighborNode.id: (Node.id, SUM(Node.costToNeighborNode, Node.costFromSource)))\}$$

$$map2: Node.id: Node \rightarrow \{list(Node.NeighborNode.id, Node.NeighborNode)\}$$

$$reduce: Node.id: \{list(prevHopFromSource, costFromSource, Node)\} \rightarrow Node.id: Node'$$
$$where$$
$$Node'.costFromSource = MIN(costFromSource),$$
$$Node'.prevHopFromSource = prevHopFromSource_{min}$$

Note: Reducer only emits value if Node structure is updated, i.e., the iteration found a new shortest path from the source.

Iteration 1:

```
Map Input: s: <s, 0, - , {(Node a, 1), (Node c, 5)}>

Map Output: (a: s, 1), (a: Node a), (c: s, 5), (c, Node c)

Reduce 1 Input: a: (s, 1, Node a)

Reduce 1 Output: a: <a, 1, s, {(Node b, 1), (Node c, 2)}>

Reduce 2 Input: c: (s, 5, Node c)

Reduce 2 Output: c: <c, 5, s, {(Node a, 3), (Node b, 9), (Node d, 2)}>
```

The reader is encouraged to continue the example and see how the solution converges in 4 iterations.

2. In which cases Map Reduce algorithm can't solve the problem efficiently? (At least two cases should be mentioned and the cause of each should be briefly explained).

   Solution.q2.a
   Solution.q2.b

3. In which cases Graph Lab is a better choice than Map Reduce to design and implement parallel systems? Why? Write down advantages and disadvantages of this framework.

   Solution.q3.a
   Solution.q3.b

4. Compare Spark and Hadoop. Write down advantages and disadvantages of each one.

   Solution.q4

CE956: Statistical Learning
Department of Computer Engineering
Sharif University of Technology
Spring 2019: Room CE204, Sat. & Mon.: 13:30-15:00

Quiz 05 (20 Points) – (May-04-2019)

Solution

Big Data: (each question 5 points)

1. What are the main characteristics of Big Data?
   - Volume (the main characteristic that makes data "big" is the large volume), Variety (multimodal data that are structured and/or unstructured), Veracity (veracity refers to the trustworthiness of the data), Velocity (velocity is the frequency of incoming data that needs to be processed).

2. What is Mapreduce? How does it work? Name a few applications of Mapreduce.
   - MapReduce is a framework for processing parallelizable problems across large datasets using a large number of nodes, collectively referred to as a cluster or a grid. Processing can occur on data stored either in a filesystem (unstructured) or in a database (structured). A MapReduce system is usually composed of three operations (or steps): (1) Map: each worker node applies the map function to the local data, and writes the output to a temporary storage. A master node ensures that only one copy of redundant input data is processed. (2) Shuffle: worker nodes redistribute data based on the output keys, produced by the map function, such that all data belonging to one key is located on the same worker node. (3) Reduce: worker nodes process each group of output data, per key, in parallel. Mapreduce can be used in many distributed processing frameworks and applications such as Hadoop and search in big data.

3. What are the pros and cons of Hadoop?
   Pros:
   - Varied Data Sources (Hadoop accepts a variety of data
   - Cost-effective (Hadoop is an economical solution as it uses a cluster of commodity hardware to store data).
   - Speed and Performance (Hadoop with its distributed processing and distributed storage architecture processes huge amounts of data with high speed).
   - Fault-Tolerant (in Hadoop 3.0 fault tolerance is provided by erasure coding).
   - Highly Available (Hadoop 3.0 HDFS supports multiple standby NameNode making the system even more highly available as it can continue functioning in case if two or more NameNodes crashes).
   - Low Network Traffic (In Hadoop, each job submitted by the user is split into a number of independent sub-tasks and these sub-tasks are assigned to the data nodes thereby moving a

small amount of code to data rather than moving huge data to code which leads to low network traffic).

- High Throughput (Hadoop stores data in a distributed fashion which allows using distributed processing with ease. A given job gets divided into small jobs which work on chunks of data in parallel thereby giving high throughput).
- Open Source (Hadoop is an open source technology).
- Scalable (Hadoop works on the principle of horizontal scalability).
- Ease of use (Hadoop framework takes care of parallel processing, MapReduce programmers does not need to care for achieving distributed processing, it is done at the backend automatically).
- Compatibility (most of the emerging technology of Big Data is compatible with Hadoop like Spark, and Flink. We use Hadoop as data storage platforms for them).
- Multiple Languages Supported (developers can code using many languages on Hadoop like C, C++, Perl, Python, Ruby, and Groovy).

Cnos:

- Issue With Small Files (Hadoop is suitable for a small number of large files but when it comes to the application which deals with a large number of small files, Hadoop fails).
- Vulnerable By Nature (Hadoop is written in Java which is a widely used programming language hence it is easily exploited by cyber criminals).
- Processing Overhead (in Hadoop, the data is read from the disk and written to the disk which makes read/write operations very expensive).
- Supports Only Batch Processing (Hadoop has a batch processing engine which is not efficient in stream processing).
- Iterative Processing (Hadoop cannot do iterative processing by itself).
- Security (Hadoop uses Kerberos authentication which is hard to manage. It is also missing encryption at storage and network levels which are a major point of concern).

4. What is Spark? When and why Spark is preferred over Hadoop?
   - Spark and its resilient distributed dataset (RDD) were developed in response to limitations in the MapReduce cluster computing paradigm, which forces a particular linear dataflow structure on distributed programs: MapReduce programs read input data from disk, map a function across the data, reduce the results of the map, and store reduction results on disk. Spark's RDDs function as a working set for distributed programs that offers a restricted form of distributed shared memory (RAM). Spark facilitates the implementation of both iterative algorithms, and interactive/exploratory data analysis. The latency of such applications may be reduced by several orders of magnitude compared to Hadoop MapReduce implementation. Spark requires a cluster manager and a distributed storage system. For cluster management, Spark supports standalone (native Spark cluster), Hadoop YARN, or Apache Mesos. For distributed storage, Spark can interface with a wide variety, including Alluxio, Hadoop Distributed File System (HDFS), MapR File System (MapR-FS), Cassandra, OpenStack Swift, Amazon S3, Kudu, or a custom solution can be implemented. Spark has Less Latency so it is relatively faster than Hadoop, since it caches most of the input data in memory by the Resilient Distributed Dataset (RDD). Spark is 100 times faster than MapReduce as everything is done here in memory. In addition, Spark supports stream processing, which involves continuous input and output of data.