Conjugate Probability Distribution and Sampling

Digya Acharya

OUTLINE:

- Questions Discussion: Probability Distribution
- Conjugate Distribution:
  - Beta-Binomial Distribution
  - Dirichlet-Multinomial Distribution
  - NormalGamma Distribution

- Sampling
  - Markov Chain
  - Markov Chain and Transition Matrix
  - Markov Chain Monte Carlo
    - Gibbs Sampling

Questions Discussion: Probability Distribution

- **Example of Bernoulli and Binomial Distribution**

- You ask a student at your school whether he or she voted for candidate A. If that person indeed voted for A, then you count it as a success. Otherwise, you count it as a failure. That's a Bernoulli experiment.

- Bernoulli trial = Bernoulli experiment

- Now you ask every person in your school whether they voted for A, and count the number of successes. That (asking all of the n students at your school) is a Binomial experiment*.
  *Assuming each person has the same probability of voting for A, and do it independently.

(Source: Quora)

Questions Discussion: Probability Distribution

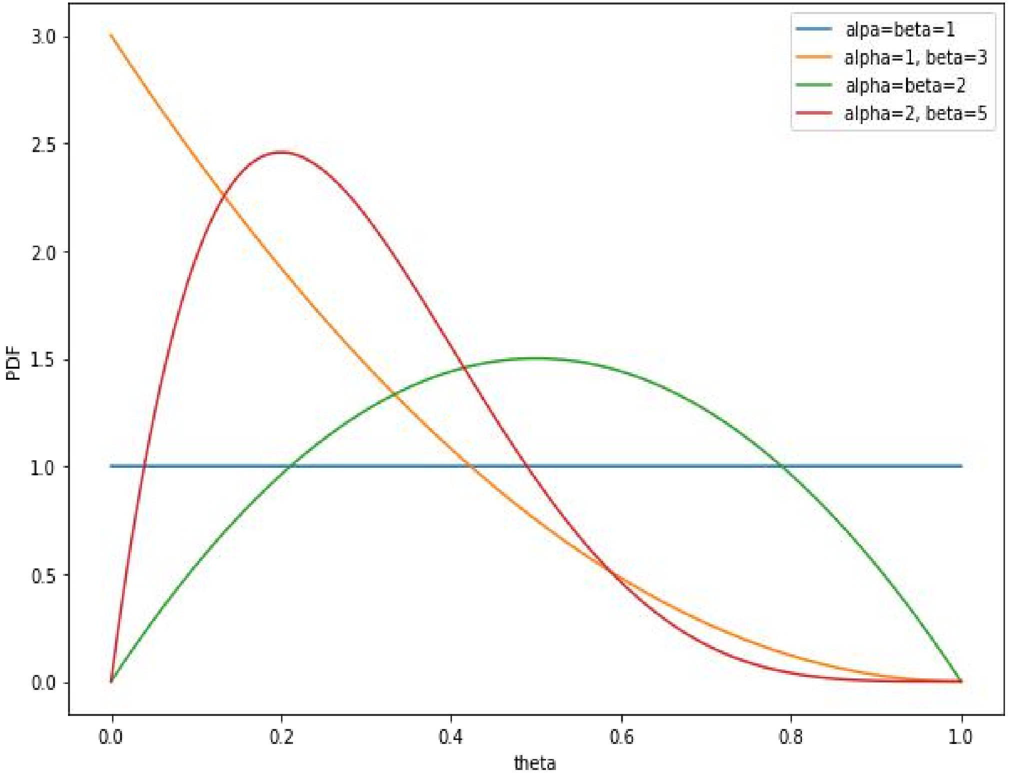


Figure 1: Pdf of beta distribution for different values of alpha and beta

Questions Discussion: Probability Distribution

- **Probability Axioms**

1. **(Nonnegativity)** $\mathrm{P(A)} \geq 0$ for every event A.
2. **(Additivity)** If $\mathrm{A1}$, $\mathrm{A2}$,... is a sequence of disjoint events then, the probability of their union satisfies:

$$\mathrm{P(A1 \cup A2 \cup \ldots) = P(A1) + P(A2) + \ldots}$$

3. **(Normalization)** The probability of the entire sample space Ω is equal to 1, i.e, $\mathrm{P}(\Omega) = 1.$

The probability distribution of pdf of ‘X’ is a function $f_X(x)$ such that for any two numbers ‘a’ and ‘b’ with $a \leq b$, we have,

$$P(a \leq X \leq b) = \int_a^b f_X(x)\,dx$$

- $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$

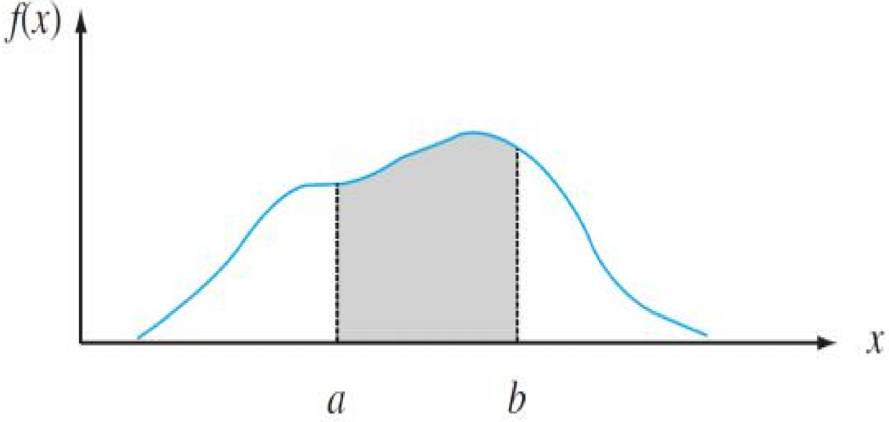
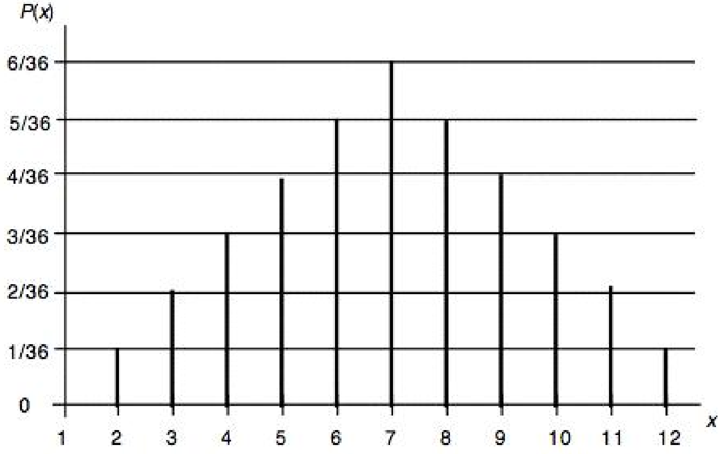


Figure 2: P(a ≤ X ≤ b) = the area under the density curve between a and b



Discrete distribution

In case of discrete distribution, PMF = Probability

In case of continuous distribution, PDF ≠ Probability

Conjugate Distribution:

If the posterior distributions $p(\theta \mid x)$ are in the same probability distribution family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions.

Example: Beta-binomial, Dirichlet-multinomial and so on.

Modelling coin toss

- **Beta-Binomial Distribution**

Suppose you visit a coin factory, pick a coin at random if α,β are the shape parameters.

You flip a coin N times and see:

$$c = (c_1, c_2)$$

i.e, c_1 heads and c_2 tails with $N = c_1 + c_2$

What is the probability of c_1 heads and c_2 tails, i.e, $p(c|\alpha, \beta)$?

Modeling typical coins from coin factory

- Using law of total probability,

$$p(c|\alpha,\beta) = \int p(c,\theta|\alpha,\beta)\,d\theta$$

$$= \int \underbrace{p(c|\theta)}_{\text{Binomial Likelihood}}\underbrace{p(\theta|\alpha,\beta)}_{\text{Beta prior}}\,d\theta$$

Modeling typical coins from coin factory

$$p(c|\alpha,\beta) = \int \frac{N!}{c1!\,c2!}\theta^{c1}(1-\theta)^{c2}\frac{\tau(\alpha+\beta)}{\tau(\alpha)\tau(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta$$

$$= \frac{N!}{c1!c2!}\frac{\tau(\alpha+\beta)}{\tau(\alpha)\tau(\beta)}\int\theta^{c1+\alpha-1}(1-\theta)^{c2+\beta-1}d\theta$$

$$= \frac{N!}{c1!c2!}\frac{\tau(\alpha+\beta)}{\tau(\alpha)\tau(\beta)}\frac{\tau(c1+\alpha)\tau(c2+\beta)}{\tau(c1+c2+\alpha+\beta)}$$

Here, the coin parameter 'θ' has been integrated out and hence no longer appears in the equation.

Using the definition of beta function,

$$p(c|\alpha,\beta) = \frac{N!}{c1!c2!}\frac{B(c1+\alpha,c2+\beta)}{B(\alpha,\beta)}$$

Binomial Posterior Distribution under Beta Prior

From Bayes rule: $p(\theta \mid c, \alpha, \beta) \; = \; \dfrac{p(c|\theta,\alpha,\beta)\; p(\theta|\alpha,\beta)}{\int_0^1 p(c|\theta,\alpha,\beta)\; p(\theta|\alpha,\beta) d\theta}$

Here, $p(c \mid \theta,\, \alpha,\, \beta)$ → Binomial Likelihood and $p(\theta|\alpha,\, \beta)$ → Beta Prior

$$p(\theta \mid c, \alpha, \beta) \; = \frac{\frac{N!}{c_1! c_2!}\; \theta^{c1}(1-\theta)^{c2}\; \frac{\tau(\alpha+\beta)}{\tau(\alpha)\tau(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 \frac{N!}{c_1! c_2!}\; \theta^{c1}(1-\theta)^{c2}\; \frac{\tau(\alpha+\beta)}{\tau(\alpha)\tau(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\, d\theta} = \frac{\theta^{c1+\alpha-1}(1-\theta)^{c2+\beta-1}}{\frac{\tau(c_1+\alpha)\tau(c_2+\beta)}{\tau(c_1+c_2+\alpha+\beta)}} =$$

$$\frac{\tau(c_1+c_2+\alpha+\beta)}{\tau(c_1+\alpha)\tau(c_2+\beta)}\; \theta^{c1\,+\,\alpha\,-\,1}(1-\theta)^{c2\,+\,\beta\,-\,1} = \frac{\tau(c_1+c_2+\alpha+\beta)}{\tau(c_1+\alpha)\tau(c_2+\beta)}\; \theta^{c1\,+\,\alpha\,-\,1}(1-\theta)^{c2\,+\,\beta\,-\,1}$$

$$= \frac{1}{B(c_1+\alpha,\; c_2+\beta)}\; \theta^{c1\,+\,\alpha\,-\,1}(1-\theta)^{c2\,+\,\beta\,-\,1}$$

$$= p(\theta|\alpha + c1, \beta + c2) \sim Beta(\alpha + c1, \beta + c2)$$

Thus, posterior of binomial (under beta) is beta with count added to the parameters.

Modelling dice roll

- **Dirichlet-Multinomial Distribution**

As in case of modelling coin toss, suppose you go to a dice factory, pick a dice at random and roll it ‘C’ times, we will need dirichlet multinomial distribution to model this.

Posterior Multinomial under Dirichlet Prior

Suppose we observe counts c = (c_1,...c_k) from a dice sampled from our factory and would like to predict the most likely parameters θ for this dice.

From Bayes rule: $p(\theta \mid c, \alpha) \;=\dfrac{p(c|\theta)\; p(\theta|\alpha)}{\int p(c|\theta)\; p(\theta|\alpha) d\theta}$

Here, $p(c \mid \theta)$ → Multinomial Likelihood and $p(\theta|\alpha)$ → Dirichlet Prior

$$p(\theta \mid c, \alpha) \;= \frac{\frac{C!}{\prod_i c_i!} \prod_i \theta_i{}^{ci} \;\frac{\tau(A)}{\prod_i \tau(\alpha_i)}\; \prod_i \theta_i{}^{\alpha i-1}}{\int p(c|\theta)\; p(\theta|\alpha) d\theta} = \frac{\frac{C!}{\prod_i c_i!}\; \frac{\tau(A)}{\prod_i \tau(\alpha_i)}\; \prod_i \theta_i{}^{\alpha i+ci-1}}{\int \frac{C!}{\prod_i c_i!}\; \frac{\tau(A)}{\prod_i \tau(\alpha_i)}\; \prod_i \theta_i{}^{\alpha i+ci-1}\, d\theta}$$

$$= \frac{\prod_i \theta_i{}^{ci+\alpha i-1}}{\frac{\prod_i \tau(ci+\alpha i)}{\tau(C+A)}} \;= \frac{\tau(C+A)}{\prod_i \tau(ci+\alpha i)} \prod_i \theta_i{}^{ci\,+\,\alpha i\,-\,1} \sim Dir(c_1 + \alpha_1, \ldots, c_k + \alpha_k)$$

So, the shape of the posterior is exactly like that of the prior with counts added to the parameters.

Normal Gamma Distribution

- conjugate prior of a normal distribution
- unknown mean and variance

Derivation: Normal Gamma Distribution

- Assume x is data point sampled using gaussian distribution then
  - $p(x|\,\mu, \lambda) = \sqrt{\frac{\lambda}{2\pi}}\, e^{\,-(1/2)\,*\,\lambda\,*\,(x-\mu)^2}$
    $= \sqrt{\frac{1}{2\pi}} * \lambda^{(1/2)} * e^{\,-(1/2)\,*\,\lambda\,*\,(\mu^2 - 2\mu x + x^2)}$
    - where μ, λ is mean and precision(inverse of sigma) of the Gaussian

- NormalGamma = $(u, \lambda \mid m, c, a, b)$
  $= N(\mu \mid m, (c\lambda)^{-1}) * Gamma(\lambda \mid a, b)$
  $= \sqrt{\frac{c\lambda}{2\pi}}\, e^{\,-(1/2)\,*\,c\lambda\,*\,(\mu-m)^2} * \frac{b^a}{\Gamma a} \lambda^{a-1}\, e^{-b\lambda}$
  $= \sqrt{\frac{c}{2\pi}}\, \frac{b^a}{\Gamma a} * \lambda^{a-(1/2)}\, e^{\,-(1/2)\,*\,\lambda\,*\,(c\mu^2 - 2mc\mu + cm^2 + 2b)}$

- Hence NormalGamma can be used as conjuagte prior for Normal distribution

- Since NormalGamma itself is a pdf:
  - $\iint \sqrt{\frac{c}{2\pi}}\, \frac{b^a}{\Gamma a} * \lambda^{a-(1/2)}\, e^{\,-(1/2)\,*\,\lambda\,*\,(c\mu^2 - 2mc\mu + cm^2 + 2b)}\, d\mu d\sigma =$
    $or, \iint \lambda^{a-(1/2)}\, e^{\,-(1/2)\,*\,\lambda\,*\,(c\mu^2 - 2mc\mu + cm^2 + 2b)}\, d\mu d\sigma = \frac{1}{\sqrt{\frac{c}{2\pi}}\, \frac{b^a}{\Gamma a}}$

Normal Distribution under Normal gamma prior

- $p(u, \lambda \mid x_{1:n}) = \frac{p(u, \lambda) * p(x_{1:n} \mid u, \lambda)}{\iint p(u, \lambda) * p(x_{1:n} \mid u, \lambda)\, d\mu\, d\sigma}$

- $p(u, \lambda) * p(x_{1:n} \mid u, \lambda)$
  $= \sqrt{\frac{c}{2\pi}} \frac{b^a}{\Gamma a} * \sqrt{\frac{1}{2\pi}} * \lambda^{a-(1/2)} e^{-(1/2) * \lambda * (c\mu^2 - 2mc\mu + cm^2 + 2b)} * \lambda^{n/2} *$
  $e^{-(1/2) * \lambda * (n\mu^2 - 2\mu \sum x_i + \sum x^2_i)}$
  $= \sqrt{\frac{c}{2\pi}} \frac{b^a}{\Gamma a} * \sqrt{\frac{1}{2\pi}} * \lambda^{(a + n/2) - (1/2)} * e^{-(1/2) * \lambda * ((c+n)\mu^2 - 2\mu (cm + \sum x_i) + cm^2 + \sum x^2_i + 2b)}$

- Now replacing:
  - $A = a + (n/2)$
    $C = c + n$
    $CM = cm + \sum x_i \;\; or \;\; M = \frac{cm + \sum x_i}{c + n}$
    $CM^2 + 2B = cm^2 + 2b + \sum x_i 2 \;\; or \;\; B = b + (1/2) * (cm^2 - CM^2 + \sum x_i^2)$

  we get,

- $p(u, \lambda) * p(x_{1:n} \mid u, \lambda) = \sqrt{\frac{c}{2\pi}} \frac{b^a}{\Gamma a} * \sqrt{\frac{1}{2\pi}} * \lambda^{A - (1/2)} * e^{-(1/2) * \lambda * (C\mu^2 - 2\mu CM + CM^2 + 2B)}$

Derivation: Posterior Normal Gamma Distribution

- $p(u, \lambda \mid x_{1:n}) = \dfrac{\sqrt{\frac{c}{2\pi}}\frac{b^a}{\Gamma a} * \sqrt{\frac{1}{2\pi}} * \lambda^{A-(1/2)} * e^{-(1/2) * \lambda * (C\mu^2 - 2\mu CM + CM^2 + 2B)}}{\iint \sqrt{\frac{c}{2\pi}}\frac{b^a}{\Gamma a} * \sqrt{\frac{1}{2\pi}} * \lambda^{A-(1/2)} * e^{-(1/2) * \lambda * (C\mu^2 - 2\mu CM + CM^2 + 2B)}\, d\mu\, d\sigma}$

  $= \dfrac{\lambda^{A-(1/2)} * e^{-(1/2) * \lambda * (C\mu^2 - 2\mu CM + CM^2 + 2B)}}{\frac{1}{\sqrt{\frac{C}{2\pi}}\frac{B^A}{\Gamma A}}}$

  $= \text{NormalGamma}(u, \lambda \mid M, C, A, B)$
- where,
  - $M = \dfrac{cm + \sum_{i=1}^{n} x_i}{c+n}$

    $C = c + n$

    $A = a + (n/2)$

    $B = b + (1/2) * (cm^2 - CM^2 + \sum_{1=1}^{n} x_i^2)$

Sampling

- Bayesians, and sometimes also frequentists, need to integrate over possibly high-dimensional probability distributions to make inference about model parameters or to make predictions.
- Bayesians need to integrate over the posterior distribution of model parameters given the data, and frequentists may need to integrate over the distribution of observables given parameter values.

Markov Chain

- The Markov property expresses the fact that at a given time step and knowing the current state, we won't get any additional information about the future by gathering information about the past.

i.e, $X_1 = PX_0$, where X_1 is the next state

P is the transition probability

X_0 is the initial state

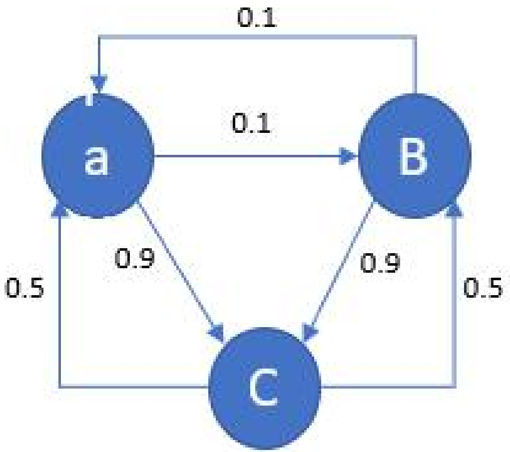
Or,

The Markov property implies that:

$$P(X_{n+1} = s_{n+1} | X_n = s_n, X_{n-1} = s_{n-1}, X_{n-2} = s_{n-2}, ...) = P(X_{n+1} = s_{n+1} | X_n = s_n)$$

Code

Markov State Diagram



$$P = \begin{bmatrix} 0 & 0.1 & 0.9 \\ 0.1 & 0.0 & 0.9 \\ 0.5 & 0.5 & 0.0 \end{bmatrix}$$

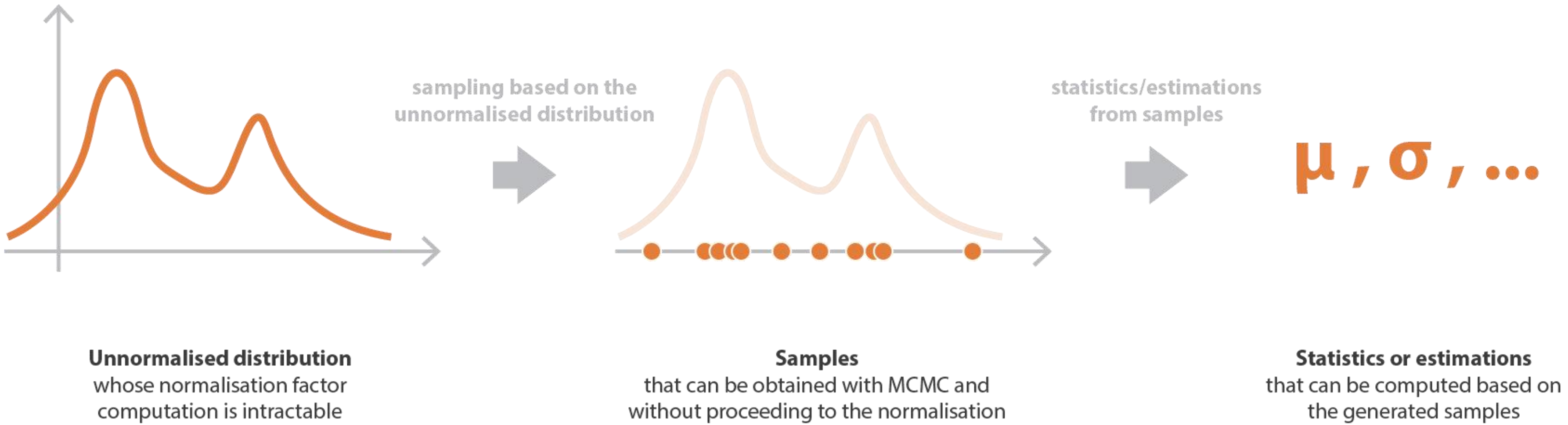
Figure 10: Markov State and the corresponding transition matrix

- From the transition matrix and initial state, we can solve markov equation until convergence.

Sampling: Markov Chain Monte Carlo

- Helps solve intractability in Bayesian Inference problem
- MCMC algorithms: Metropolis-Hasting and Gibbs Sampling
- Normalization factor becomes intractable in higher dimension
- Sampling approach: generates samples from a given probability distribution
- Named “Markov chain Monte Carlo” because we obtain samples using markov chain method

Estimate parameters using samples



Picture credit: Joseph Rocca

Sample states using MCMC

Picture credit: Joseph Rocca

Gibbs Sampling

- MCMC algorithm
- Assumes that the conditional distribution can be computed even if the joint probability is intractable.
- Conditional distribution converges to joint distribution on limit.

$$P(X_i | X_1, \ldots, X_{i-1}, X_{i+1}, \ldots X_n)$$

$$= \frac{P(X_1, \ldots X_n)}{P(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots X_n)} \ \alpha \ P(X_1, \ldots X_n)$$

[Code]

Key Takeaways

- We discussed some questions from probability distribution.
- We discussed about conjugate distribution: Beta-Binomial, Dirichlet-Multinomial, Normal Gamma
- We discussed about the basics of sampling
  - Markov Chain
  - Markov Chain and Transition Matrix
  - Markov Chain Monte Carlo
    - Gibbs Sampling

References:

- Manandhar Suresh. *Bayesian ML : Posterior Distributions and Mixture Models Continuous Probability Density Function*
- Bishop Christopher. *Pattern Recognition and Machine Learning*

Thank you