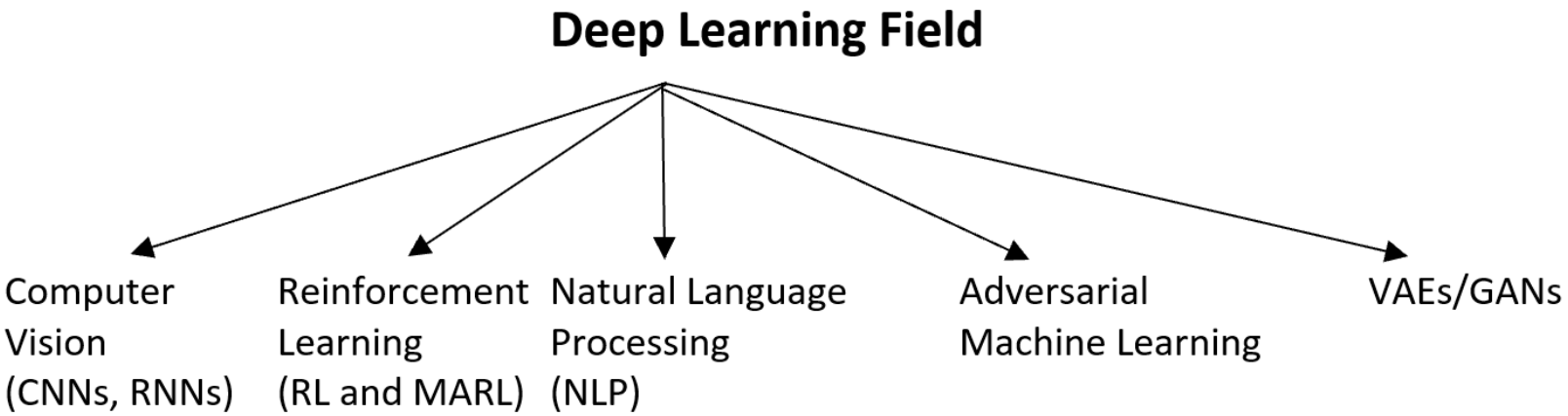
CPEG 589 – Advanced Deep Learning
Lecture 1

Outline

- Introduction and Course Overview
- Review of PDF, Joint PDF, Mutual Information, Conditional Probability, Expectation, Marginalization, Bayes Theorem
- Popular Probability Distributions in Machine Learning
- Logistic Regression
- Naïve Bayes Classification
- Maximum Likelihood Estimation and Maximum A Posteriori Estimation

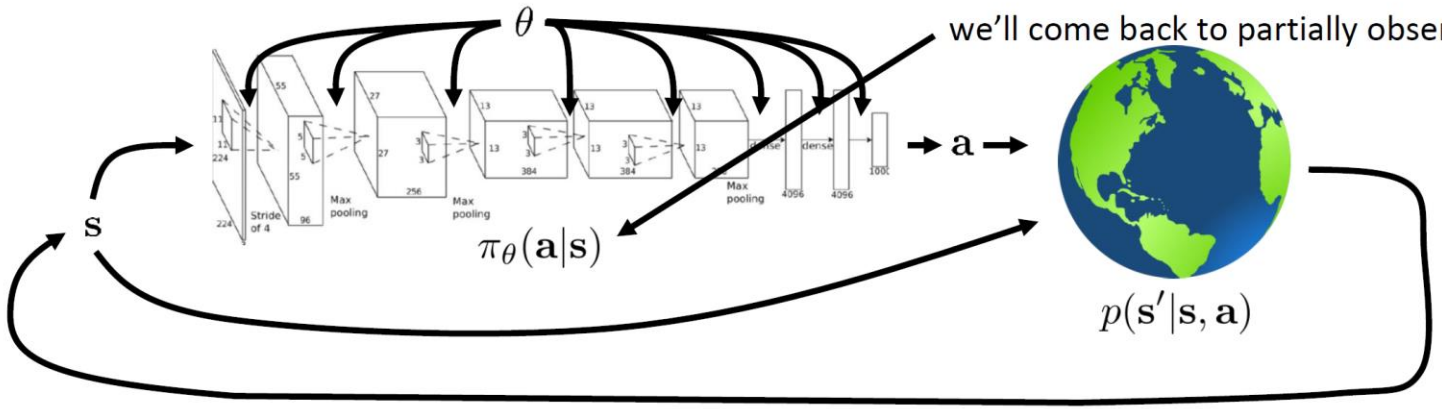
Course Topics

- Goal : Cover latest Developments in AI and Deep Learning
- Be able to understand nicely the latest research papers in top AI Conferences
- Challenge : Difficult Mathematics and Programming



Challenge

The goal of reinforcement learning



$$\underbrace{p_\theta(\mathbf{s}_1, \mathbf{a}_1, \ldots, \mathbf{s}_T, \mathbf{a}_T)}_{\pi_\theta(\tau)} = p(\mathbf{s}_1) \prod_{t=1}^{T} \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$$

$$\theta^\star = \arg\max_\theta E_{\tau \sim p_\theta(\tau)} \left[\sum_t r(\mathbf{s}_t, \mathbf{a}_t) \right]$$

RL

Direct policy differentiation

$$\theta^\star = \arg\max_\theta J(\theta)$$

$$J(\theta) = E_{\tau\sim\pi_\theta(\tau)}[r(\tau)]$$

$$\nabla_\theta J(\theta) = E_{\tau\sim\pi_\theta(\tau)}[\nabla_\theta \log \pi_\theta(\tau) r(\tau)]$$

log of both sides

$$\underbrace{\pi_\theta(\mathbf{s}_1, \mathbf{a}_1, \ldots, \mathbf{s}_T, \mathbf{a}_T)}_{\pi_\theta(\tau)} = p(\mathbf{s}_1)\prod_{t=1}^{T}\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$$

$$\log \pi_\theta(\tau) = \log p(\mathbf{s}_1) + \sum_{t=1}^{T}\log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) + \log p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$$

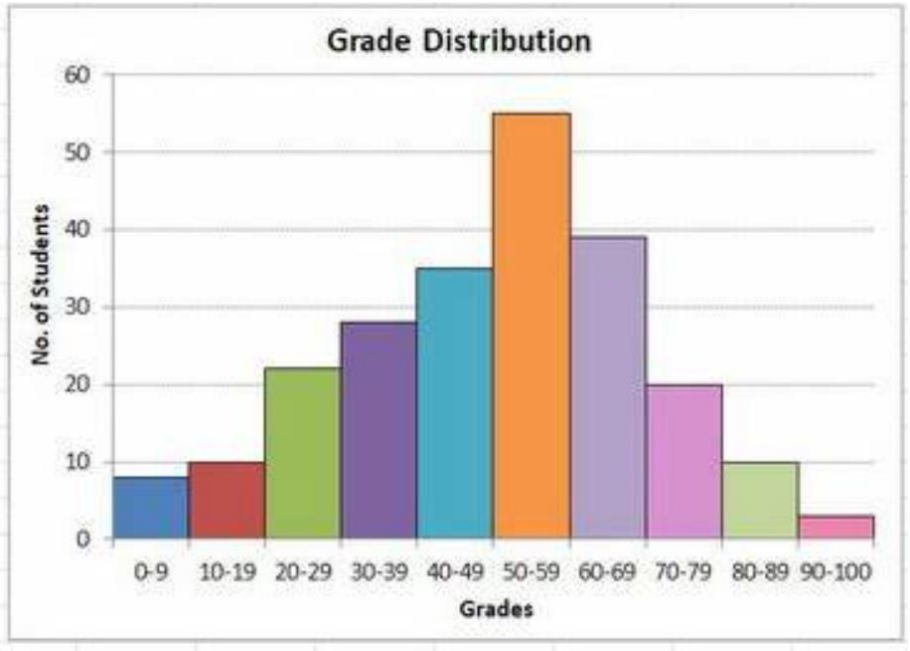
$$\nabla_\theta \left[\log p(\mathbf{s}_1) + \sum_{t=1}^{T}\log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) + \log p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)\right]$$

$$\nabla_\theta J(\theta) = E_{\tau\sim\pi_\theta(\tau)}\left[\left(\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\right)\left(\sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t)\right)\right]$$

Probability Mass and Density Functions

- PDF – continuous data f(x)
- PMF – discrete data P(X)

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



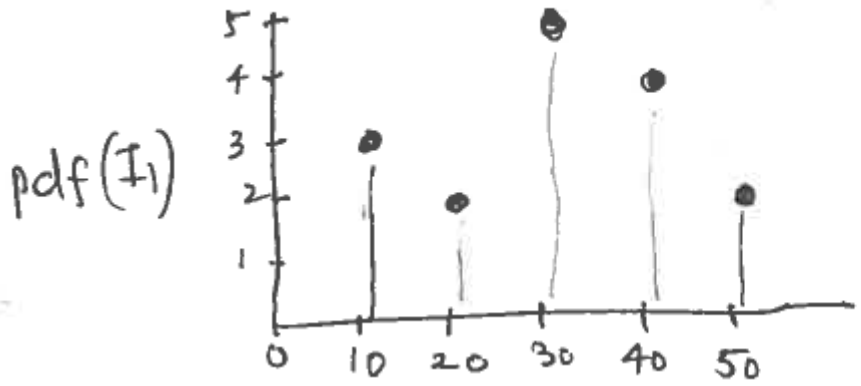
PDF Eample

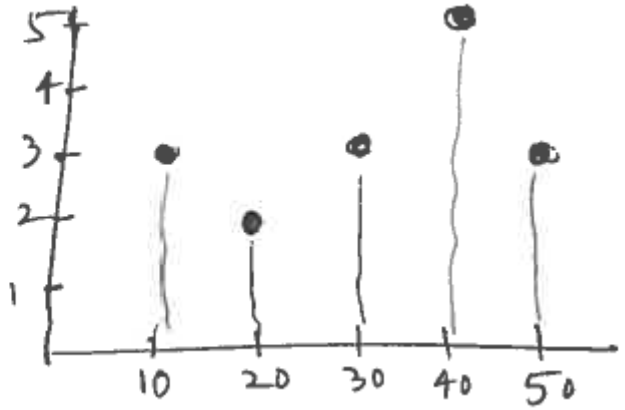
Problem #2:

a) (10 points) Suppose two 4x4 gray scale images are given to you.

$$I1 = \begin{bmatrix} 10 & 20 & 30 & 40 \\ 50 & 30 & 50 & 40 \\ 40 & 30 & 30 & 10 \\ 10 & 20 & 30 & 40 \end{bmatrix} \qquad I2 = \begin{bmatrix} 10 & 20 & 40 & 40 \\ 10 & 50 & 50 & 40 \\ 40 & 30 & 30 & 50 \\ 10 & 20 & 30 & 40 \end{bmatrix}$$

Show the histograms for pdf(I_1) and pdf(I_2) (Hint: X axis in pdf will have values of 10,20,30,40,50)





Joint PDF

b) (10 points) Show the joint pdf(I_1,I_2) for the images in part a)

$$pdf(a,\tilde{b}) = \frac{1}{M.N}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}\delta(a, I_1(x,y)) \,.\, \delta(\tilde{b},\ \tilde{I_2}(x,y))$$

$$I1 = \begin{bmatrix} 10 & 20 & 30 & 40 \\ 50 & 30 & 50 & 40 \\ 40 & 30 & 30 & 10 \\ 10 & 20 & 30 & 40 \end{bmatrix} \quad I2 = \begin{bmatrix} 10 & 20 & 40 & 40 \\ 10 & 50 & 50 & 40 \\ 40 & 30 & 30 & 50 \\ 10 & 20 & 30 & 40 \end{bmatrix}$$

pdf (10,10) = 2
(10,20) = 0
(10,30) = 0
(10,40) = 0
(10,50) = 1

pdf (20,10) = 0
(20,20) = 2
(20,30) = 0
(20,40) = 0
(20,50) = 0

pdf (30,10) = 0
(30,20) = 0
(30,30) = 3
(30,40) = 1
(30,50) = 1

pdf (40,10) = 0
(40,20) = 0
(40,30) = 0
(40,40) = 4
(40,50) = 0

pdf (50,10) = 1
(50,20) = 0
(50,30) = 0
(50,40) = 0
(50,50) = 1

Mutual Information

c) (10 points) Compute the Mutual Information MI between the two images I_1 and I_2.

$$MI(I_1, \tilde{I}_2) = \sum_{a=0}^{255} \sum_{b=0}^{255} p(a, \tilde{b}) . \log \left(\frac{p(a, \tilde{b})}{p(a) . p(\tilde{b})} \right)$$

$$2 \log\left(\frac{2}{3\times3}\right) + 1\times\log\left(\frac{1}{3\times3}\right) + 2\times\log\left(\frac{2}{2\times2}\right)$$

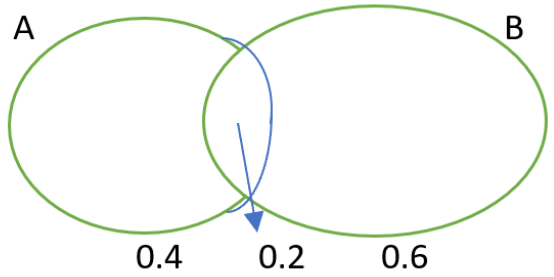
$$+ 3 \log\left(\frac{3}{5\times3}\right) + 1\times\log\left(\frac{1}{5\times5}\right) + 1\times\log\left(\frac{1}{5\times3}\right)$$

$$+ 4 \log\left(\frac{4}{4\times5}\right) + 1\times\log\left(\frac{1}{2\times3}\right) + 1\times\log\left(\frac{1}{2\times3}\right)$$

Relationship between Joint and Conditional Probability

P(A,B) = P(A|B) P(B)

Or P(A,B) = P(B|A) P(A)



$$P(A|B) = \frac{P(A,B)}{P(B)} = 0.2/0.6 = 1/3$$

$$P(B|A) = \frac{P(A,B)}{P(A)} = 0.2/0.4 = 1/2$$

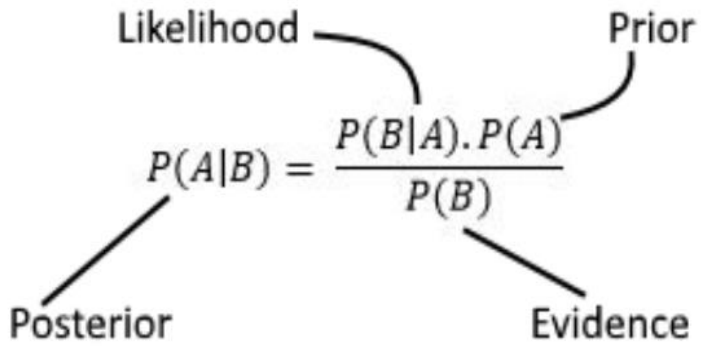
Baye's Rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = ((1/2) \times 0.4)/0.6 = 1/3$$

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = ((1/3) \times 0.6)/0.4 = 1/2$$

Bayes Rule - Example

- Ex1 : Probability a card is 7, given it is red
- P(7|R) = P(R|7) P(7)/P(R) =
- ((1/2) x (1/13))/(1/2) = 1/13



- EX2 : It rains 3 days out of 30, 6 days out of 30 are cloudy, 60% of the time it starts out cloudy – it rains. What is the probability it will rain, given it is cloudy:
- P(rain|cloudy) = P(cloudy|rain) P(rain)/P(cloudy)
- P(rain|cloudy) = (0.6 x 3/30)/(6/30) = 0.3

Chain Rule of Joint Probability

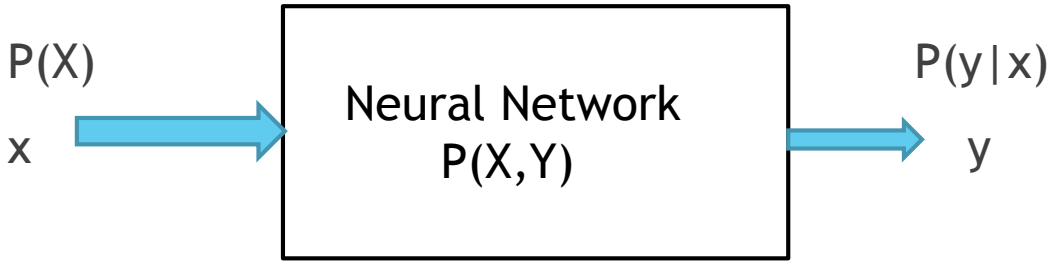
- P(A,B) = P(A|B)P(B) = P(B|A)P(A)
- P(A,B,C) = P(A|B,C)P(B,C) = P(A|B,C)P(B|C)P(C)
- P(A,B,C|D) = P(C|D) P(A,B|C,D)
- For independent variables:
- P(A,B,C) = P(A) P(B) P(C)

General case

$$P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i | x_1 \ldots x_{i-1})$$

Neural Network – Probability View

- When we train a Neural Network, we can think of it as learning the joint PDF P(x,y)



$$P(y|x) = \frac{P(x|y)\,P(Y)}{P(X)} = \frac{P(X,Y)}{P(X)}$$

Marginal Distribution

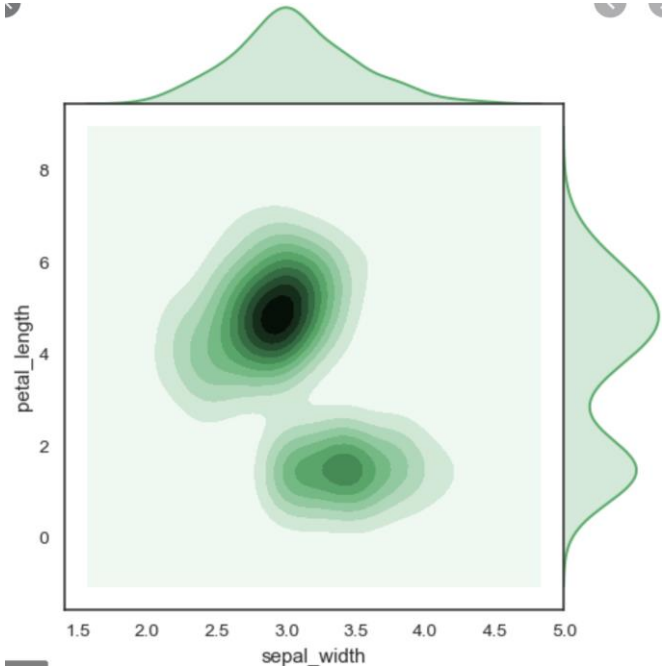
- $f_X(x) = \int f(x,y)dy$
- $f_Y(y) = \int f(x,y)dx$
- For Discrete case:

$$f_X(x) = \sum_y f_{XY}(x,y)$$

and

$$f_Y(y) = \sum_x f_{XY}(x,y)$$

- For marginalization from conditional distribution:
- $P(X) = \sum_y P(X|Y)P(Y)$
- $P(Y) = \sum_x P(Y|X)P(X)$



Expectation

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

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{XY}(x, y)\, dy dx$$

$$\mu = E(x) = \sum xp(x)$$

Conditional Expectation:

$$E\left[X \middle| Y = y\right] = \sum_{x} x p_{X|Y}\left(x \middle| y\right)$$

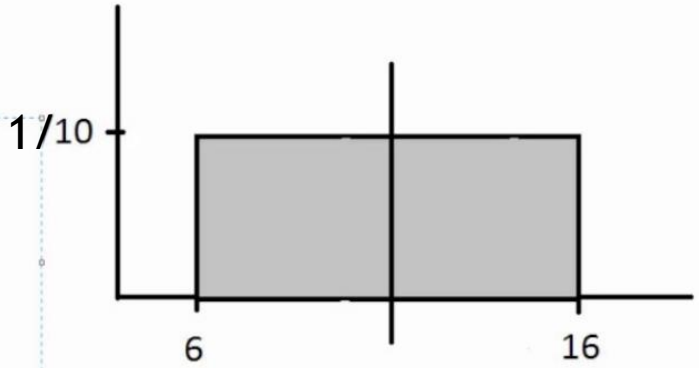
Popular Probability Distributions

Uniform Distribution

The time it takes me to wash the dishes is uniformly distributed between 6 minutes and 16 minutes.

Uniform: $X \sim U(a, b)$ where $a < x < b$

- $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$
- $P(c < X < d) = (d - c)\left(\frac{1}{b-a}\right)$
- mean $\mu = \frac{a+b}{2}$
- standard deviation $\sigma = \sqrt{\frac{(b-a)^2}{12}}$



The standard uniform density has parameters a = 0 and b = 1, the PDF for standard uniform density is given by:

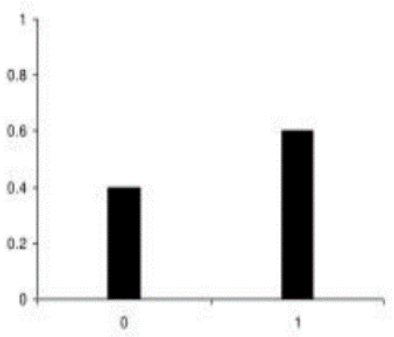
$$f(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & otherwise \end{cases}$$

Bernoulli Distribution

- It has only two possible outcomes, 1 (success) and 0 (failure)

- Example: Coin Toss - Probability of getting a head = 0.5,
  Probability of getting a tail = 0.5
- If it is unfair coin, then P(head) = 0.4, P(tail) = 0.6
- The probability mass function is given by: $p^x (1-p)^{1-x}$

$$p(x) = P[X = x] = \begin{cases} q = 1 - p & x = 0 \\ p & x = 1 \end{cases}$$



- E[X] = 1xP + 0x(1-P) = P
- Variance = $E[X^2] - E[X]^2 = P - P^2 = P(1-P)$

$$X = \begin{cases} 1 & \text{Bernoulli trial} = \mathbf{S} \\ 0 & \text{Bernoulli trial} = \mathbf{F} \end{cases}$$

Binomial Distribution

- An experiment with two possible outcomes repeated n times

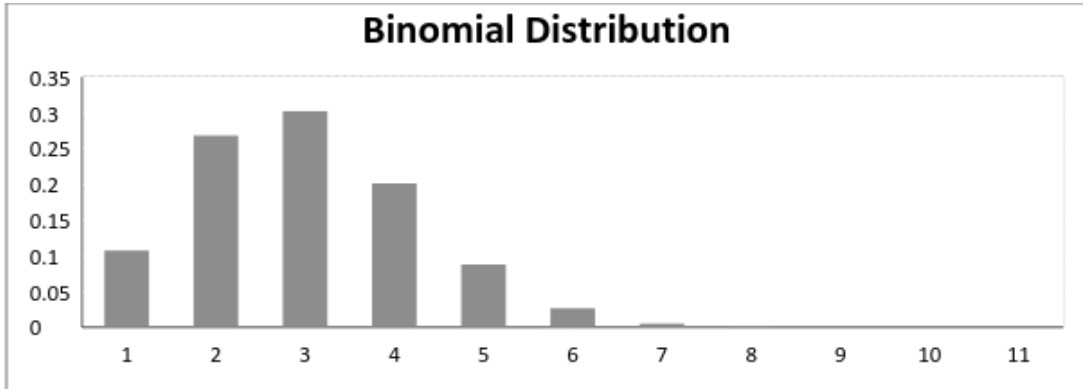
$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!\,x!} p^x q^{n-x}$$

where
n = the number of trials (or the number being sampled)
x = the number of successes desired
p = probability of getting a success in one trial
$q = 1 - p$ = the probability of getting a failure in one trial

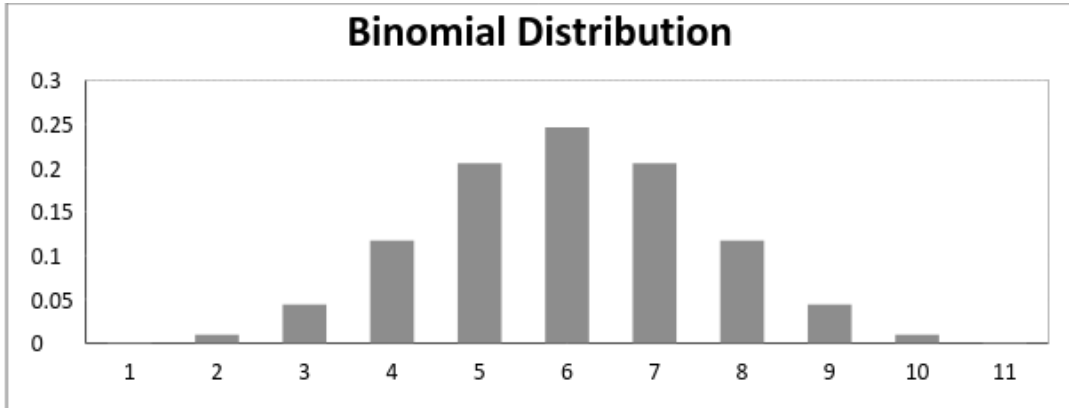
- Mean $\mu = n*p$
- Variance = $n*p*q$
- Example: P(John makes 2 out 3 free throws) = ${}^3C_2 \times (0.9)^2 \times (0.1)^1 = 0.243$
- Bernoulli Distribution is a special case of Binomial Distribution with a single trial

Binomial Distribution

- A binomial distribution graph where the probability of success does not equal the probability of failure looks like



- With p(success) = p(failure)



Gaussian or Normal Distribution

- Gaussian or normal distribution, 1D

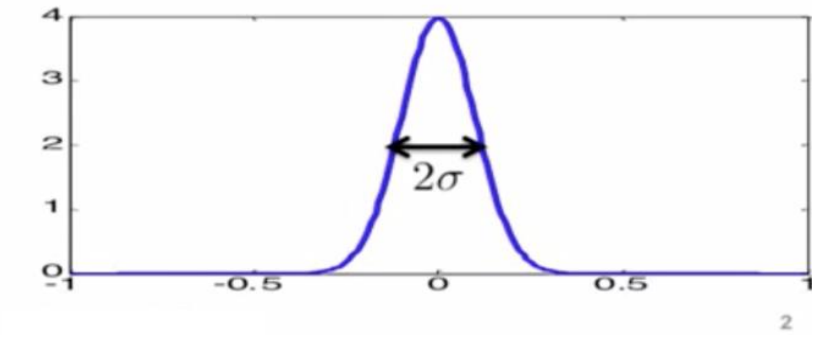
$$\mathcal{N}(x\ ;\ \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}(x-\mu)^2/\sigma^2\right]$$

  - Parameters: mean μ, variance σ^2 (standard deviation σ)

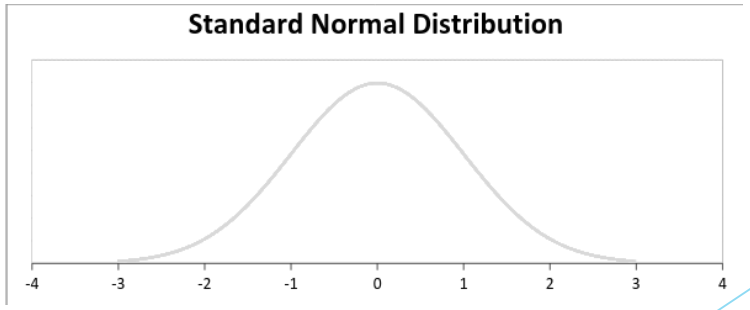
Maximum Likelihood estimates

$$\hat{\mu} = \frac{1}{N}\sum_i x^{(i)}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_i (x^{(i)} - \hat{\mu})^2$$



- Standard Normal Distribution has 0 mean and unit variance



Poisson Distribution

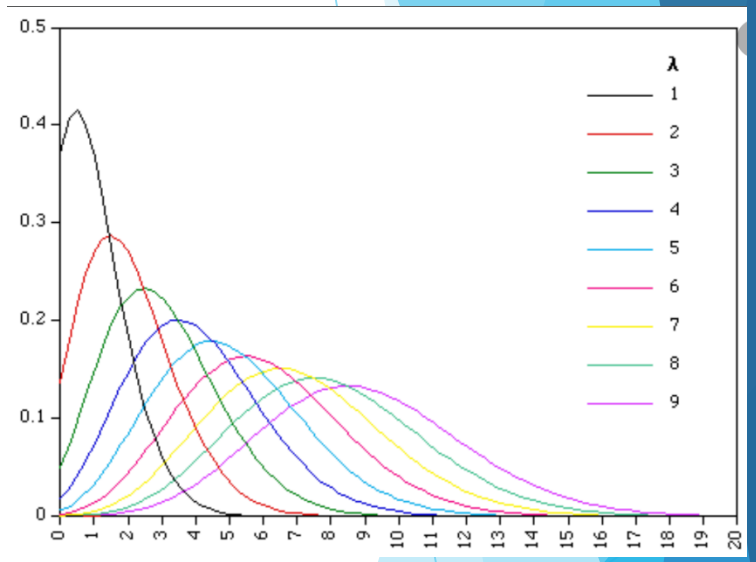
Probability of events for a Poisson distribution

An event can occur 0, 1, 2, … times in an interval. The average number of events in an interval is designated λ (lambda). λ is the event rate, also called the rate parameter. The probability of observing k events in an interval is given by the equation

$$P(k \text{ events in interval}) = e^{-\lambda}\frac{\lambda^k}{k!}$$

where

- λ is the average number of events per interval
- e is the number 2.71828... (Euler's number) the base of the natural logarithms
- k takes values 0, 1, 2, …
- $k! = k \times (k-1) \times (k-2) \times \ldots \times 2 \times 1$ is the factorial of k.



The average number of homes sold is 2 homes per day. What is the probability that exactly 3 homes will be sold tomorrow?

Solution: This is a Poisson experiment in which we know the following:

•μ = 2; since 2 homes are sold per day, on average.,x = 3; since we want to find the likelihood that 3 homes will be sold tomorrow.

$$P(x; \mu) = (e^{-\mu})(\mu^x) / x! = P(3; 2) = (2.71828^{-2})(2^3) / 3! = 0.180$$

Exponential Distribution

- Used in modeling device life time
  - The probability density function of the exponential distribution is given by

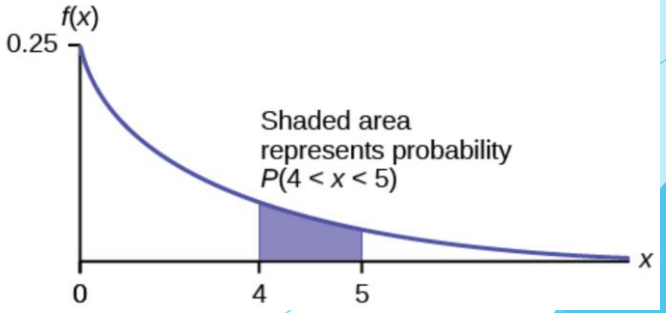
$$f(x) = \lambda e^{-\lambda x} \qquad x > 0, \lambda > 0$$

$$f(x) = \frac{1}{\mu} e^{-\frac{1}{\mu}x}$$

  - E[X] = 1/λ
  - λ = 1/μ
  - Var(X) = 1/λ²

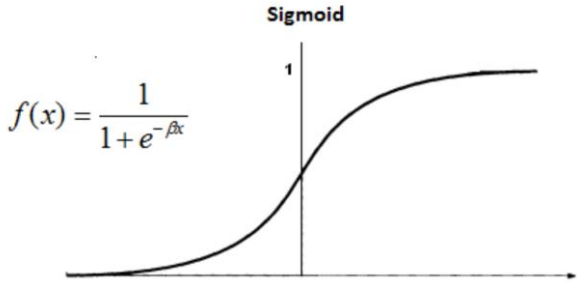
$$F(x) = \int_0^x f(x)dx = 1 - e^{-\lambda x} \qquad x > 0, \lambda > 0$$

- Example: A teller spends 4 minutes on average with a customer. What is the probability he/she will spend 4-5 minutes = 0.6321



Logistic Regression

- Used in Binary classification
- Suppose x_{i1} , x_{i2} ,.. x_{in} are some features. We want to create a linear binary classifier
- $h(x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots\ldots + \beta_n x_{in}$
- $x_i^T = [1 \;\; x_{i1} \;\; x_{i2} \;\; \ldots x_{in}]$ $\beta^T = [\beta_0 \;\; \beta_1 \;\; \ldots. \;\; \beta_n]$
- $h(x_i) = \beta^T x_i$
- To convert our hypothesis to a probability, we can pass it through sigmoid function
- $h(x_i) = \sigma(\beta^T x_i)$



- Suppose y_i is the expected output of our network, corresponding to input x_i, we can define conditional probabilities on the two labels as:
- $P(y_i = 1 \mid x_i, \beta) = h(x_i)$
- $P(y_i = 0 \mid x_i, \beta) = 1 - h(x^i)$

Logistic Regression

- $P(y_i = 1 \mid x_i, \beta) = h(x_i)$
- $P(y_i = 0 \mid x_i, \beta) = 1 - h(x^i)$
- Combine into one equation as:
- $P(y_i = 0 \mid x_i, \beta) = (h(x_i))^{yi} (1 - h(x^i))^{(1- yi)}$
- For m training examples, we want to maximize the probability of classification (maximum likelihood):
- $\beta^* = \underset{\beta}{\operatorname{argmax}} \prod_{i=1}^{m} (h(xi))^{yi} (1 - h(xi))^{(1- yi)} = \underset{\beta}{\operatorname{argmax}}(L(\beta)$
- (find β that maximizes the likelihood of correctly predicting the data)
- $L(\beta) = \prod_{i=1}^{m} (h(xi))^{yi} (1 - h(xi))^{(1-yi)}$ (multiplication of probabilities can becomes 0 for large m)
- Thus instead, we try to maximize the log likelihood, i.e.,
- $L = \text{Log}(L(\beta)) = \sum_{i=}^{m} yi \log(h(xi) + (1 - yi)\log(1 - h(xi)$ maximization can be comnverted tp minimization by multiplying the loss with negative sign
- $L(\beta) = - \sum_{i=}^{m} yi \log(h(xi) + (1 - yi)\log(1 - h(xi)$

Logistic Regression

- $\beta^{*} = argmin_{\beta} J(\beta) = argmin_{\beta}[-\sum_{i=}^{m} yi \log(h(xi) + (1 - yi)\log(1 - h(xi)]$
- The above is a transcendental equation and cannot be solved in closed form
- Use gradient descent to solve it (we initialize β randomly and repeated apply)
- $\beta = \beta - \alpha(\nabla_{\beta}(J(\beta))$
- Once the above iteration converges, the β obtained can be plugged into our model to determine the probability of x_i as:
- $h(x_i) = \sigma(\beta^T x_i)$
- How to determine $(\nabla_{\beta}(J(\beta))$?

$$L(\beta) = -\sum_{i=1}^{m} yi \log(h(xi)) + (1 - y_i) \log(1 - h(xi))$$

$$h(xi) = \frac{1}{1 + e^{-\beta^T xi}}$$

$$L(\beta) = -\sum_{i=1}^{m} \underbrace{yi\,(0 - \log(1+e^{-\beta^T xi}))}_{(1)} + \underbrace{(1-yi)\log\left(1 - \left(\frac{1}{1+e^{-\beta^T xi}}\right)\right)}_{(2)}$$

Logistic Regression

$$L(\beta) = -\sum_{i=1}^{m} y_i \log(h(x_i)) + (1-y_i)\log(1-h(x_i)) \quad (4)$$

$$h(x_i) = \frac{1}{1+e^{-\beta^T x_i}}$$

$$L(\beta) = -\sum_{i=1}^{m} \underbrace{y_i\left(0-\log\left(1+e^{-\beta^T x_i}\right)\right)}_{(1)} + \underbrace{(1-y_i)\log\left(1-\left(\frac{1}{1+e^{-\beta^T x_i}}\right)\right)}_{(2)} \quad (10)$$

Focusing on the partial derivative w.r.t. β for equation (10) for part (1).

$$\frac{\partial(1)}{\partial\beta} = \frac{y_i\, x_i\, e^{-\beta^T x_i}}{1+e^{-\beta^T x_i}}$$

Note that $\frac{e^{-\beta^T x_i}}{1+e^{-\beta^T x_i}} = -\left(\frac{1}{1+e^{-\beta^T x_i}} - 1\right) = -h(x_i)+1 = 1-h(x_i)$

$$\frac{\partial(1)}{\partial\beta} = y_i\, x_i (1-h(x_i)) \quad (11)$$

$$\frac{\partial(2)}{\partial\beta} = \frac{(1-y_i)}{1-\frac{1}{1+e^{-\beta^T x_i}}} \times \frac{\partial}{\partial\beta}\left(1-\frac{1}{1+e^{-\beta^T x_i}}\right) = \frac{1-y_i}{1-h(x_i)}\left[\frac{x_i\, e^{-\beta^T x_i}}{\left(1+e^{-\beta^T x_i}\right)^2}\right]$$

Logistic Regression

$$\frac{\partial (2)}{\partial \beta} = \frac{1-y_i}{1-h(x_i)} \times x_i \left(h(x_i)-1\right) h(x_i) \quad \text{---(12)}$$

$$\frac{\partial L(\beta)}{\partial \beta} = -\left(\frac{\partial (1)}{\partial \beta} + \frac{\partial (2)}{\partial \beta}\right)$$

$$= -\left[y_i x_i \left(1-h(x_i)\right) + \frac{(1-y_i)}{1-h(x_i)} \times x_i \left(h(x_i)-1\right) h(x_i)\right]$$

$$= -\left[y_i x_i - y_i x_i h(x_i) - (1-y_i) x_i h(x_i)\right]$$

$$= -\left[y_i x_i - y_i x_i h(x_i) - x_i h(x_i) + y_i x_i h(x_i)\right]$$

$$= -\left(y_i - h(x_i)\right) x_i$$

for all training samples m

$$\boxed{\frac{\partial L(\beta)}{\partial \beta} = -\sum_{i=1}^{m} (y_i - a_i)\, x_i}$$

where $a_i = h(x_i)$ = actual output
y_i = expected output

Naïve Bayes

- A simple technique for constructing classifiers
- Assumes that the value of a particular feature is independent of the value of any other feature, given the class variable
- An advantage of naive Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification

$$p(C_k \mid x_1, \ldots, x_n)$$

for each of K possible outcomes or *classes* C_k

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})} \qquad \text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

$$p(C_k, x_1, \ldots, x_n)$$

which can be rewritten as follows, using the chain rule for repeated applications of the definition of conditional probability:

$$\begin{aligned} p(C_k, x_1, \ldots, x_n) &= p(x_1, \ldots, x_n, C_k) \\ &= p(x_1 \mid x_2, \ldots, x_n, C_k)\, p(x_2, \ldots, x_n, C_k) \\ &= p(x_1 \mid x_2, \ldots, x_n, C_k)\, p(x_2 \mid x_3, \ldots, x_n, C_k)\, p(x_3, \ldots, x_n, C_k) \\ &= \cdots \\ &= p(x_1 \mid x_2, \ldots, x_n, C_k)\, p(x_2 \mid x_3, \ldots, x_n, C_k) \cdots p(x_{n-1} \mid x_n, C_k)\, p(x_n \mid C_k)\, p(C_k) \end{aligned}$$

Naïve Bayes

$$p(C_k, x_1, \ldots, x_n) = p(x_1 \mid x_2, \ldots, x_n, C_k)\, p(x_2 \mid x_3, \ldots, x_n, C_k) \cdots p(x_{n-1} \mid x_n, C_k)\, p(x_n \mid C_k)\, p(C_k)$$

- Since our assumption is that features are independent (naïve)

$$\begin{aligned} p(C_k \mid x_1, \ldots, x_n) &\propto p(C_k, x_1, \ldots, x_n) \\ &\propto p(C_k)\, p(x_1 \mid C_k)\, p(x_2 \mid C_k)\, p(x_3 \mid C_k) \cdots \\ &\propto p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k), \end{aligned}$$

$$p(C_k \mid x_1, \ldots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)$$

where the evidence $Z = p(\mathbf{x}) = \sum_k p(C_k)\, p(\mathbf{x} \mid C_k)$ is a scaling factor dependent only on $x_1, \ldots, x_n$,

$$\hat{y} = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}}\ p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k).$$