Data Analytics
EEE 4774 & 6777

Module 2

Frequentist vs. Bayesian Probability

Spring 2022

Uncertainty & Probability

- ***Uncertainty*** in data:
  - inherent in the observed physical process (e.g., voltage measurement in power grid, # customers in a market)
  - noise in measurement (e.g., hardware/software limitations)
  - finite data size (i.e., lack of access to the entire population)

- ***Probability***:
  - a consistent framework for quantification and manipulation of uncertainty
  - helps in decision making (e.g., our brains)

Probability & Statistics in Data Science

- Compute statistics such as mean and variance to get insight
  - E.g., mean and standard deviation of age or height data in this class
- Build probabilistic models of data to use statistics in a systematic way
  - Classify data instances
  - Predict future values
  - Estimate missing values
  - Generate realistic (simulated) data
  - Assess the confidence in decisions
    - class probabilities, confidence intervals for predictions

dog 0.93 prob.
cat 0.05
duck 0.02

house value $500k ± 20k with a prob. 0.95
(480k, 520k)

Frequentist probability

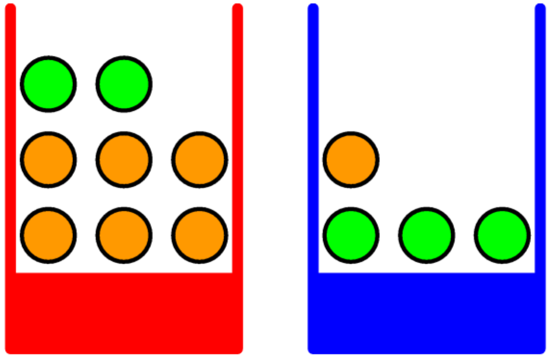
- Frequency of observations

color of bag

  - *marginal probability* $p(Y = o) = n_o/n$

color of ball

  - *joint probability* $p(X = r, Y = g) = n_{rg}/n$
  - *conditional probability* $p(Y = g|X = b) = n_{bg}/n_b$



- Sum rule: $p(X) = \sum_Y p(X,Y)$

$$p(X=b) = p(X=b, Y=o) + p(X=b, Y=g)$$

- Product rule: $p(X,Y) = p(Y|X)p(X)$

$$p(X=b, Y=g) = P(Y=g|X=b)P(X=b)$$

		$X = r$ (n_r)	$X = b$ (n_b)	
Y	g	$n_{rg} = 2$	$n_{bg} = 3$	n_g
	o	$n_{ro} = 6$	$n_{bo} = 1$	n_o

$$n = n_r + n_b = 12$$
$$= n_g + n_o$$

- Bayes Theorem: $p(X|Y) = \dfrac{p(X,Y)}{p(Y)} = \dfrac{p(Y|X)p(X)}{\sum_X p(Y|X)p(X)}$

$p(X = b|Y = g) = ?$

$$= \frac{P(Y=g|X=b)\,p(X=b)}{P(Y=g)}$$

$$P(Y=g) = P(Y=g|X=b)\,p(X=b) + P(Y=g|X=r)P(X=r)$$

$$= \frac{\frac{3}{4}\cdot\cancel{\frac{1}{2}}}{\frac{3}{4}\cdot\cancel{\frac{1}{2}} + \frac{2}{8}\cdot\cancel{\frac{1}{2}}} = \frac{3}{4}$$

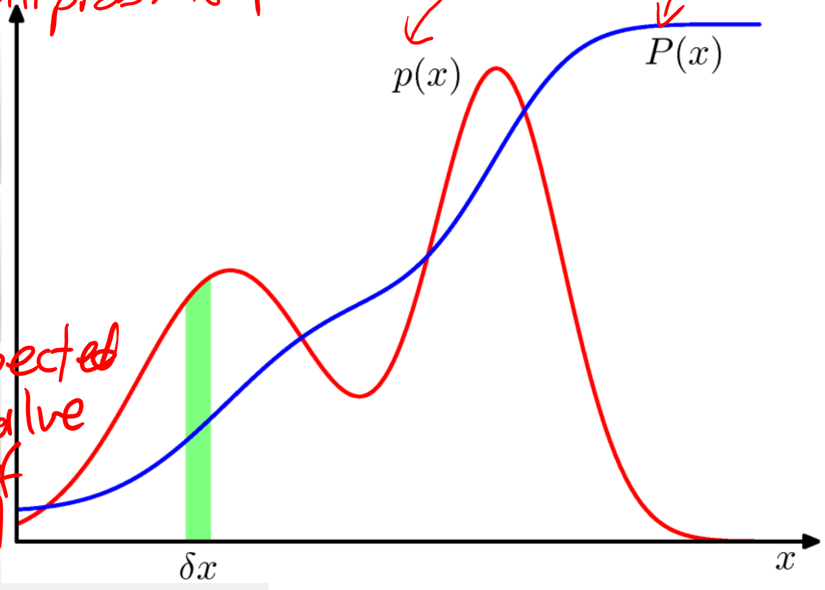
Probability

- $p(x)$: Probability density/mass function of a continuous/discrete variable

pmf pdf cdf

sum of all prob. is 1

$p(x \in (a,b)) = \int_a^b p(x)\,dx$	$p(x) \geq 0$	$\int_{-\infty}^{\infty} p(x)\,dx = 1$
cdf $P(y) = \int_{-\infty}^{y} p(x)dx$	$p(\mathbf{x}) \geq 0$ vector = $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$	$\iiint p(\mathbf{x})\,d\mathbf{x} = 1$
$p(x) = \int p(x,y)\,dy$ sum rules	$p(x,y) = p(y\|x)p(x)$ product rule	$E[f(x)] = \sum_x f(x)p(x)$ expected value of f(x)
$E[f(x)] = \int f(x)p(x)dx$ continuous x	$E_x[f(x,y)] = \int f(x,y)p(x,y)dx$	$E_{x\|y}[f(x,y)\|y] = \int f(x,y)p(x\|y)dx$
$Var[f]$ $= E[(f(x) - E[f(x)])^2]$ $= E[f(x)^2] - E[f(x)]^2$	$Cov[x,y]$ $= E_{x,y}[(x - E[x])(y - E[y])]$ $= E_{x,y}[xy] - E[x]E[y]$	$Cov[\mathbf{x},\mathbf{y}]$ $= E_{\mathbf{x},\mathbf{y}}[(\mathbf{x} - E[\mathbf{x}])(\mathbf{y} - E[\mathbf{y}])^T]$ $= E_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^T] - E[\mathbf{x}]E[\mathbf{y}^T]$



$$\begin{array}{cc} & y \\ x & \begin{bmatrix} & \\ & \end{bmatrix} \end{array}$$

Bayesian Probability

- ***Classical/Frequentist*** interpretation of probability ~ frequencies of ***repeatable*** events

- ***Bayesian probability*** ~ a quantification of uncertainty
  - repeatable and ***non-repeatable*** events, e.g., the probability of *a dragon flying through the window*
  - ***update with evidence***, e.g., it is shown that there exist dragons in Florida, and there are small ones that can fit through a window.

0.3

$$p(\mathbf{x}|D) = \frac{p(D|\mathbf{x})\, p(\mathbf{x})}{p(D)}$$

10^{-20}

10^{-18}

$\simeq$ 10^{-2}

posterior α *likelihood* x *prior*

Bayesian Probability

$$p(\mathbf{x}|D) = \frac{p(D|\mathbf{x})\, p(\mathbf{x})}{p(D)}$$

posterior α *likelihood* x *prior*

indep. of x

$\max_x p(x|D) \equiv \max_x p(D|x)p(x)$

x : Class label

$x \in \{\text{dog}, \text{cat}\}$

$\hat{x} = \arg\max_x p(D|x)p(x)$

image

- ***Prior probability*** is not an arbitrary choice, *reflects common sense* (or uninformative)

0.75 dog

0.25 cat

- ***Challenge:*** for predictions and model comparison, marginalization typically difficult!

$$p(D) = \iiint p(D|\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x}$$

Bayesian vs. Frequentist

	Bayesian	Frequentist
Likelihood	fixed data, random parameters	random data, fixed parameters
Model Selection	training data: evidence (Occam's razor)	training + validation data: cross validation (may be computationally cumbersome)
Regularization	naturally provided by ***prior*** (prevents overfitting)	needs additional penalty
Accuracy	naturally provided by ***posterior*** (quality evaluation)	needs additional techniques (confidence interval, bootstrap)

- Bayesian ***prior*** *may not be realistic*, but *more and more training data* decreases the effect of prior

- Advances in *computational power*, as well as *techniques for computing posterior & marginal*

(e.g., *sampling techniques* such as MCMC, and *approximate inference* such as variational Bayes)

promote Bayesian approach, enable its use in *Big Datasets*.