Intro to Machine Learning (CS436/CS580L)

Lecture 3: Probability & Statistics

Xi Peng, Fall 2018

Thanks to Tom Mitchell, Andrew Ng, Ben Taskar, Carlos Guestrin, Eric Xing, Hal Daume III, David Sontag, Jerry Zhu, Tina Eliassi-Rad, and Chao Chen for some slides & teaching material.

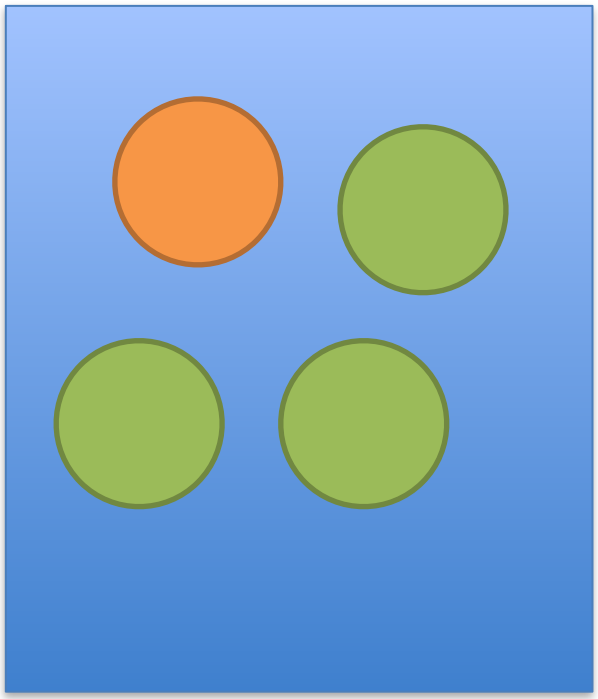
This Class

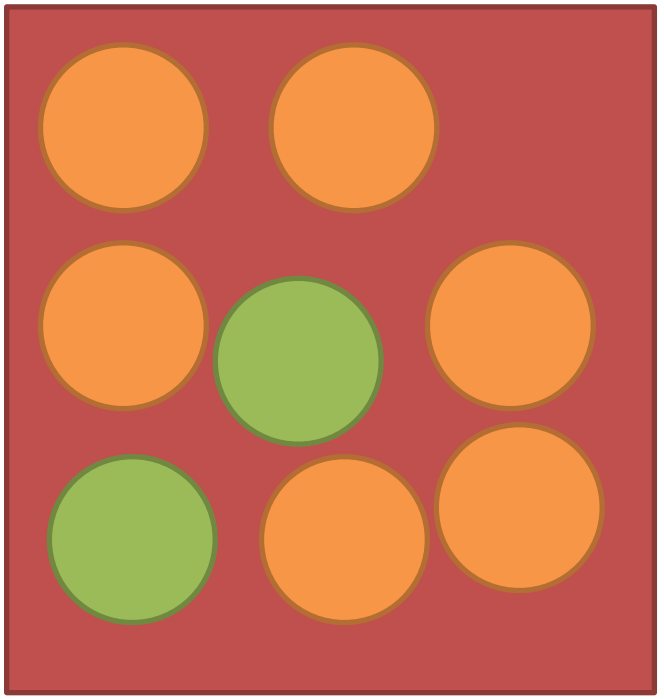
- Probability and Statistics
  - Basic probabilities
  - Naïve Bayes Classification

Median

Boxes and Balls

- 2 Boxes, one red and one blue.
- Each contain colored balls.





Boxes and Balls

- Suppose we randomly select a box, then randomly draw a ball from that box.
- The identity of the Box is a **random variable**, *B*.
- The identity of the ball is a **random variable**, *L*.
- *B* can take 2 values, *r,* or *b*
- *L* can take 2 values, *g* or *o*.

Boxes and Balls

- Given some information about *B* and *L*, we want to ask questions about the likelihood of different events.
- What is the probability of selecting a green ball?
- If I chose an orange ball, what is the probability that I chose from the blue box?

Some basics

- The **probability** of an event is the fraction of times that the event occurs out of *n* trials, as *n* approaches infinity.
- Probabilities lie in the range [0,1]
- **Mutually exclusive** events are events that cannot simultaneously occur.
  - The sum of the likelihoods of all mutually exclusive events must equal 1.

Joint Probability – P(X,Y)

- A Joint Probability function defines the likelihood of two (or more) events occurring.

	Orange	Green	
Blue box	1	3	4
Red box	6	2	8
	7	5	12

- Let n_{ij} be the number of times event *i* and event *j* simultaneously occur.

$$p(X = x_i, Y = y_i) = \frac{n_{ij}}{N}$$

Generalizing the Joint Probability

	n_{ij}		$r_i = \sum_j n_{ij}$
	$c_j = \sum_i n_{ij}$		$\sum_i \sum_j n_{ij} = N$

Marginalization

- Consider the probability of X irrespective of Y.

$$p(X = x_j) = \frac{c_j}{N}$$

- The number of instances in column j is the sum of instances in each cell

$$c_j = \sum_{i=1}^{L} n_{ij}$$

- Therefore, we can **marginalize** or “sum over” Y :

$$p(X = x_j) = \sum_{i=1}^{L} p(X = x_j, Y = y_i)$$

Conditional Probability

- Consider only instances where X = x_j.
- The fraction of these instances where Y = y_i is the conditional probability
  - "The probability of y given x"

$$p(Y = y_i | X = x_j) = \frac{n_{ij}}{c_j}$$

  - Why?

Relating the Joint, Conditional and Marginal

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$

$$= p(Y = y_j | X = x_i) p(X = x_i)$$

- Similarly ……
- Bayes' rule

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

- p(B=r | L=o)

Sum and Product Rules

- In general, we'll refer to a distribution over a random variable as p(X) and a distribution evaluated at a particular value as p(x).
- Addition rule:
  - A and B mutually exclusive: p(A or B) = p(A) + p(B)
    - Example: throwing a dice **once**
  - Otherwise: p(A or B) = p(A) + p(B) – p(A and B)
    - Example: getting an A in ML / in Data Structure

Sum and Product Rules

- Multiplication rule:
  - A and B independent: p(A and B) = p(A)p(B)
    - Example: throwing a dice **twice**
  - Otherwise: p(A and B) = p(B|A)p(A) = p(A|B)p(B)
    - Example: pick the orange ball from the blue box
- Chain rule:
  - p(A, B, C, D) = p(D | C,B,A) p(C | B,A) p(B | A) p(A)
- If two events are **independent** then,

  p(X, Y) = p(X)p(Y) → p(X|Y) = p(X)
- **Conditional independent**

  p(X, Y | Z) = p(X | Z) p(Y | Z)

(Conditional) Independence

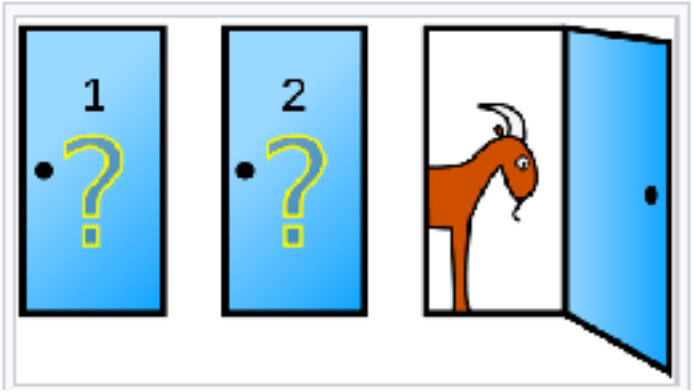
- If two events are **independent** then,

  $p(X, Y) = p(X)p(Y) \rightarrow p(X|Y) = p(X)$

- **Conditional independent**
  $p(X, Y \mid Z) = p(X \mid Z)\, p(Y \mid Z)$
- Examples:
  - height and vocabulary are not independent. But they are conditioned on age.
  - Two dice are independent, but they are not conditioned on their sum.

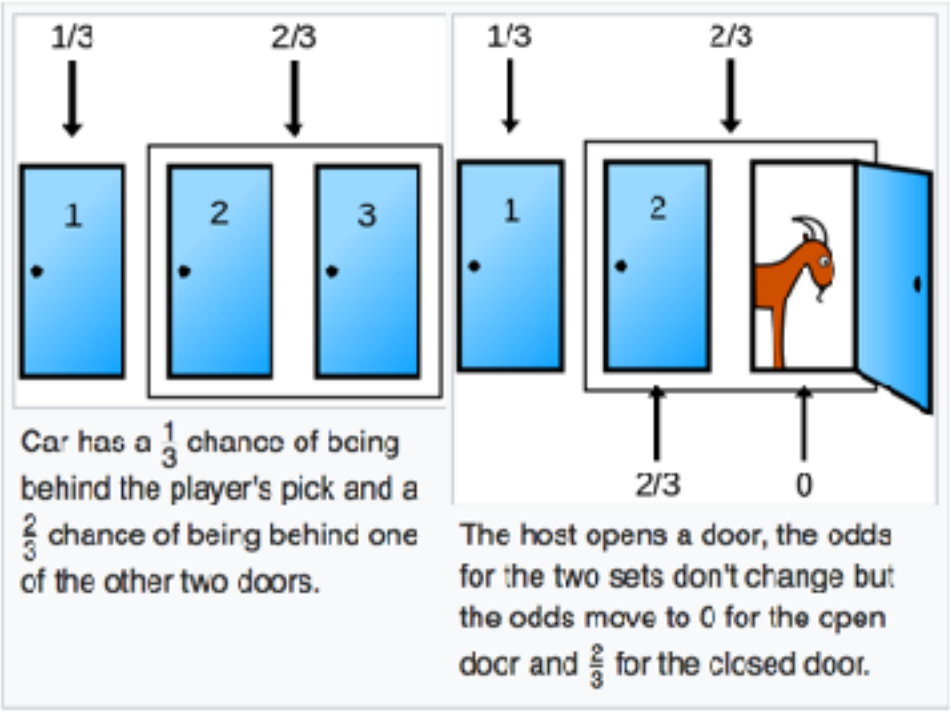
The Monty Hall Problem

- Game show
- Three doors,
  - behind one of them is a fancy race car,
  - behind each of the other two is a goat
- If you first choose one door (not opening)
- The host pick one of the other two with a goat behind it
- Should you switch?



The Monty Hall Problem

- p(car=2) = p(car=2 or 3) * p(car != 3 | car = 2 or 3)
  = 2/3 * 1



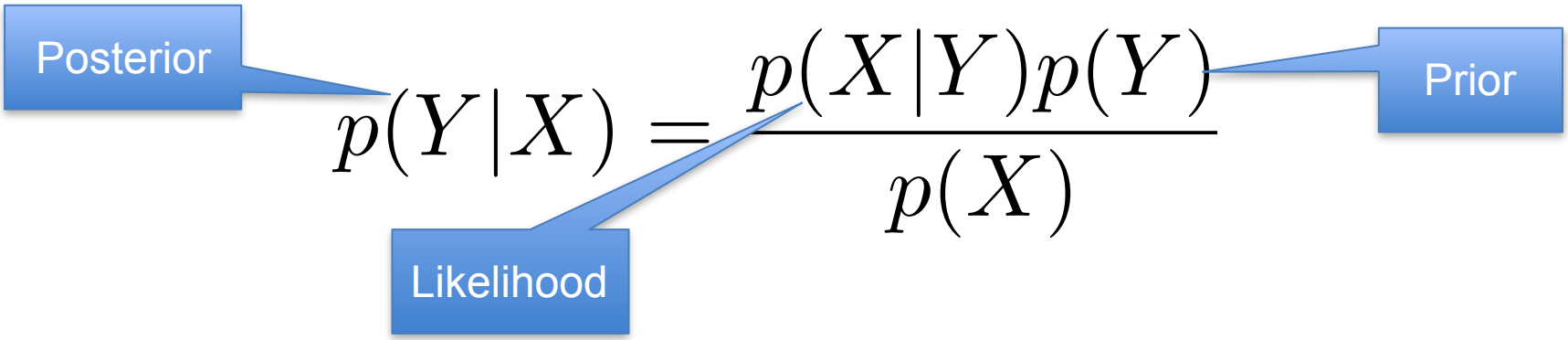
Behind door 1	Behind door 2	Behind door 3	Result if staying at door #1	Result if switching to the door offered
Car	Goat	Goat	**Wins car**	Wins goat
Goat	**Car**	Goat	Wins goat	**Wins car**
Goat	Goat	**Car**	Wins goat	**Wins car**

In Classification

- Testing: observation/features → class
  p(B=r | L=o)
- Training:
  - But in high dimension, exponentially many possible L values, cannot assume enough observation conditioned on each L
  - However, classes are limited: class → observation
- How to related them?

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

Interpretation of Bayes Rule



- **Observation**: X
- **Prediction:** y = class of X
- **Prior**: Information we have before observation.
- **Posterior**: The distribution of Y after observing X
- **Likelihood:** The likelihood of observing X given Y
- **What about p(X)?**
- **Sometimes p(X) can be ignored**

Naïve Bayes Classification

- This is a simple case of a simple classification approach.
- Here the Box is the class, and the colored ball is a feature, or the observation.
- We can extend this Bayesian classification approach to incorporate more **independent** features.

Naïve Bayes Example Data

X1	X2	X3	C
HOT	LIGHT	SOFT	RED
COLD	HEAVY	SOFT	RED
HOT	HEAVY	FIRM	RED
HOT	LIGHT	FIRM	RED
COLD	LIGHT	SOFT	BLUE
COLD	HEAVY	FIRM	BLUE
HOT	HEAVY	FIRM	BLUE
HOT	LIGHT	FIRM	BLUE
HOT	HEAVY	FIRM	?????

Naïve Bayes Classification

- Some theory first

$$c^* = \underset{c}{\operatorname{argmax}}\, p(c|x_1, x_2, \ldots, x_n)$$

$$c^* = \underset{c}{\operatorname{argmax}}\, \frac{p(x_1, x_2, \ldots, x_n|c)p(c)}{p(x_1, x_2, \ldots, x_n)}$$

- Naïve conditional independence assumption

$$p(x_1, x_2, \ldots, x_n|c) = p(x_1|c)p(x_2|c)\cdots p(x_n|c)$$

Naïve Bayes Classification

- Assuming conditional independent features simplifies the math.

$$c^* = \underset{c}{\operatorname{argmax}} \frac{p(x_1|c)p(x_2|c)\cdots p(x_n|c)p(c)}{p(x_1, x_2, \ldots, x_n)}$$

$$c^* = \underset{c}{\operatorname{argmax}}\, p(x_1|c)p(x_2|c)\cdots p(x_n|c)p(c)$$

- Training, testing

Naïve Bayes Example Data

HOT	LIGHT	SOFT	RED
COLD	HEAVY	SOFT	RED
HOT	HEAVY	FIRM	RED
HOT	LIGHT	FIRM	RED
COLD	LIGHT	SOFT	BLUE
COLD	HEAVY	FIRM	BLUE
HOT	HEAVY	FIRM	BLUE
HOT	LIGHT	FIRM	BLUE
HOT	HEAVY	FIRM	?????

$$c^* = \underset{c}{\operatorname{argmax}}\, p(x_1|c)p(x_2|c)\cdots p(x_n|c)p(c)$$

Naïve Bayes Example Data

Prior:

Likelihood:

Posterior:

Continuous Probabilities

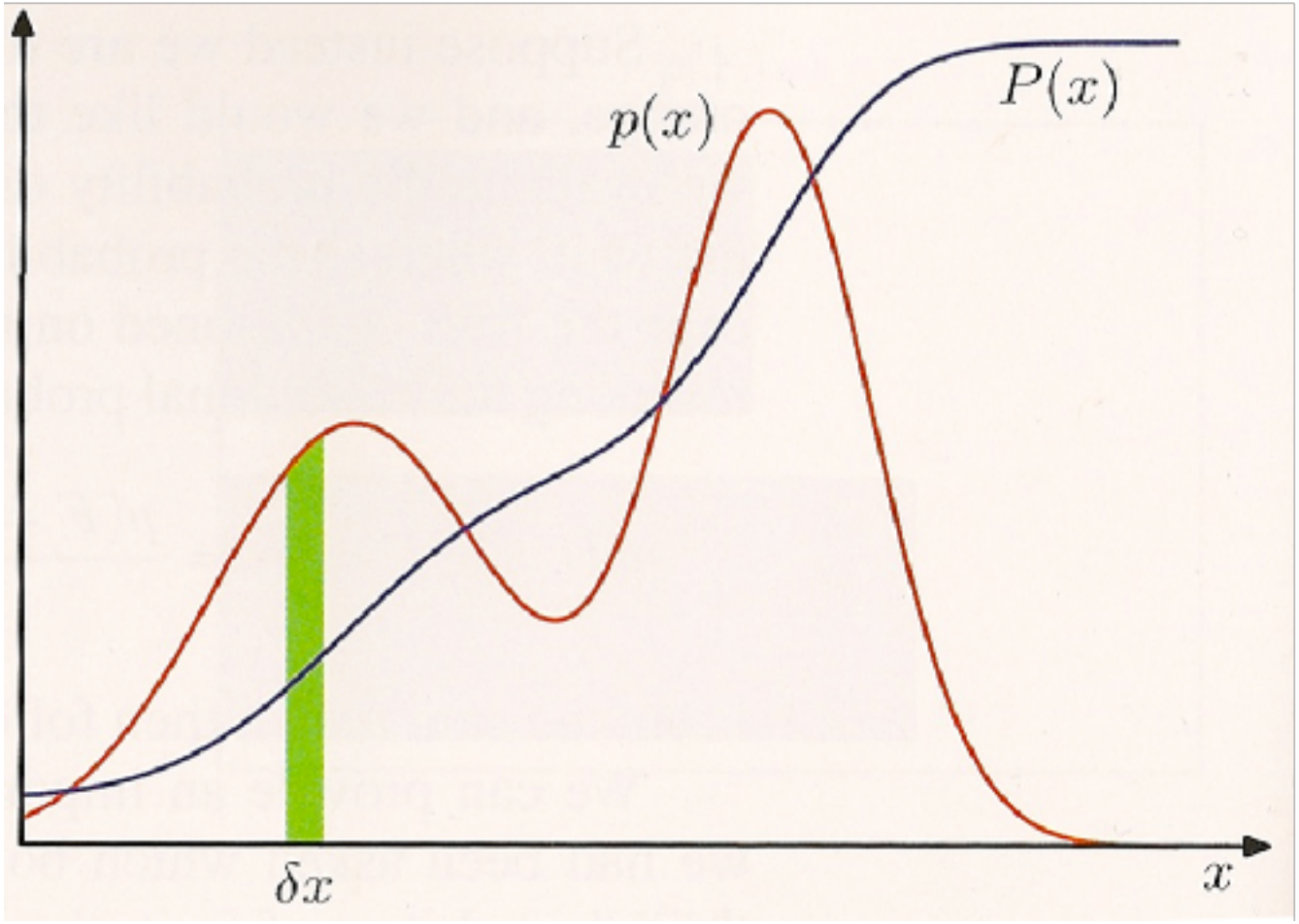
- So far, X has been **discrete** where it can take one of M values.
- What if X is continuous?
- Now p(x) is a continuous **probability density function**.
- The probability that x will lie in an interval (a,b) is:

$$p(x \in (a,b)) = \int_b^a p(x)dx$$

- Cumulative distribution function (CDF)

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x \qquad P'(x) = p(x).$$

Continuous probability example



Properties of probability density functions

$$p(x) \geq 0$$

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

$$p(x) = \int p(x,y)\,\mathrm{d}y$$

$$p(x,y) = p(y|x)p(x).$$

- Addition/multiplication/chain rule
- Bayes' rule
  - Replace sum with integral

Expected Values (mean, average)

- Given a random variable, with a distribution p(X), what is the expected value of X?

$$\mathbb{E}[x] = \sum_x p(x)x$$

$$\mathbb{E}[x] = \int p(x)x\mathrm{d}x$$

Variance

- The **variance** of a random variable describes how much variability around the expected value there is.
- Calculated as the expected squared error.

$$\mathrm{var}[f] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2]$$

$$\mathrm{var}[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

- Why?
- Standard deviation

Covariance

- The covariance of two random variables expresses how they vary together.

$$\begin{aligned}\mathrm{cov}[x, y] &= \mathbb{E}_{x,y}[(x - \mathbb{E}(x))(y - \mathbb{E}[y])] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

- If two variables are **independent**, their covariance equals zero.
- cov[x,x]?

Intuition

- Expectation
  - The expected value of a function is the **hypothesis**
- Variance
  - The variance is the **confidence** in that hypothesis

Gaussian Distribution

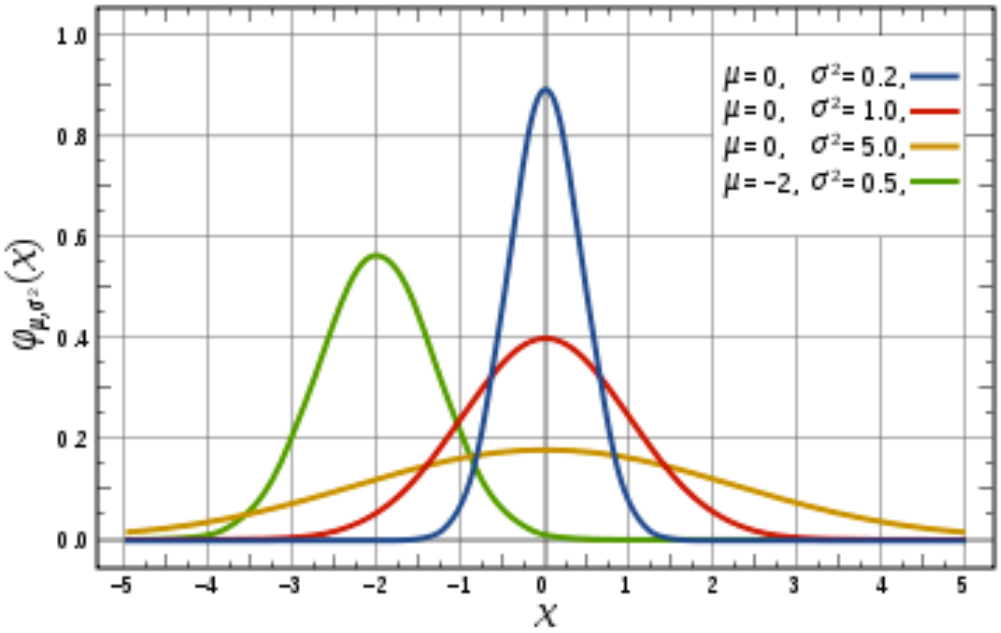
- One Dimension

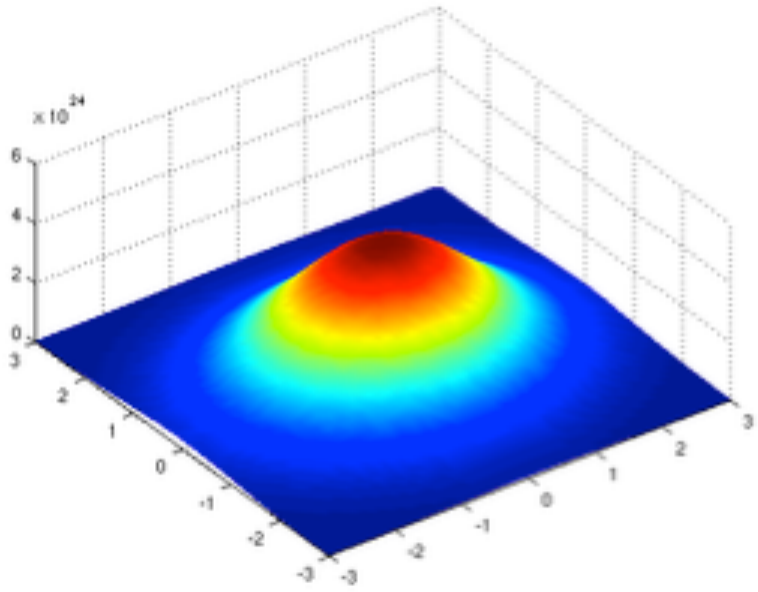
$$N(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

- D-Dimensions
  - (will return after Linear Algebra review)

$$N(x|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

Gaussians





To Do

- Review “Bishop”: Ch 1 & 2
- Review Slides L3 & L4