SDU Summer School

Deep Learning

Summer 2022

Welcome to the Summer School

Statistics

- Why Probability?
- Introduction to Probability
- Random Variables & Distributions
- Important Distribution Functions

Why Probability?

- Much of CS deals with entities that are certain
  - CPU executes flawlessly
    - At least almost … there are CPU bugs and CPUs can also be broken
  - CS and software engineers work in clean and certain environment
  - Surprising that ML heavily uses probability theory

- Reasons for ML use of probability theory
  - Must always deal with uncertain quantities
    - Also with non-deterministic (stochastic) quantities
  - Many sources for uncertainty and stochasticity

Sources of Uncertainty

1. Inherent stochasticity of system being modeled
   - Subatomic particles are probabilistic
   - Cards shuffled in random order

2. Incomplete observability
   - Deterministic systems appear stochastic when not all variables are observed

3. Incomplete modeling
   - Discarded information results in uncertain predictions

Practical to use uncertain rule

- Simple rule “Most birds fly” is cheap to develop and broadly useful
- Rules of the form “Birds fly, except for very young birds that have not learned to fly, sick or injured birds that have lost ability to fly, flightless species of birds...” are expensive to develop, maintain and communicate
    - Also still brittle and prone to failure

Tools of Probability

- Probability theory was originally developed to analyze frequencies of events
  - Such as drawing a hand of cards in poker
  - These events are repeatable
  - If we repeated experiment infinitely many times, proportion of p of outcomes would result in that outcome

- Is it applicable to propositions not repeatable?
  - Patient has 40% chance of flu
    - Cannot make infinite replicas of the patient
  - We use probability to represent degree of belief

- Former is frequentist probability, latter Bayesian, just fyi.

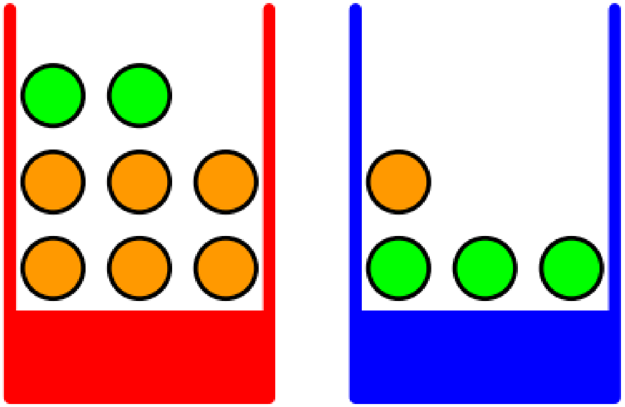
Statistics

- Why Probability?
- Introduction to Probability
- Random Variables & Distributions
- Important Distribution Functions

Definition of Probability

- To begin with, we shall define the probability of an event to be the fraction of times that event occurs out of the total number of trials, in the limit that the total number of trials goes to infinity.

- Let's look at the following simple Experiment:
  - Two boxes, red (r) and blue (b)
  - Each boxes contain either apples (a) or oranges (o)
  - The box that will be chosen is a random variable B. It takes the values r or b
  - The fruit sampled out of the selected box is donated F and takes the values a and o



➢ Taken from Christopher M. Bishop, "Pattern Recognition and Machine Learning"

Random Variables

- Let's assume we select the blue box 60% of the time
- Repeating the box selection infinite number of times, we say that the probability of selecting the red box is 4/10 and the blue box 6/10.
- We formally write this as follows:

$$p(B = r) = \frac{4}{10} \qquad p(B = b) = \frac{6}{10}$$

  - Note that, by definition, probabilities must lie in the interval [0, 1]
  - Also, if the events are mutually exclusive and if they include all possible outcomes, then we see that the probabilities for those events must sum to one.

➢ Taken from Christopher M. Bishop, "Pattern Recognition and Machine Learning"

Joint Probability Distributions

- We can now ask questions such as:
  - “what is the overall probability that the selection procedure will pick an apple?”
  - “given that we have chosen an orange, what is the probability that the box we chose was the blue one?”

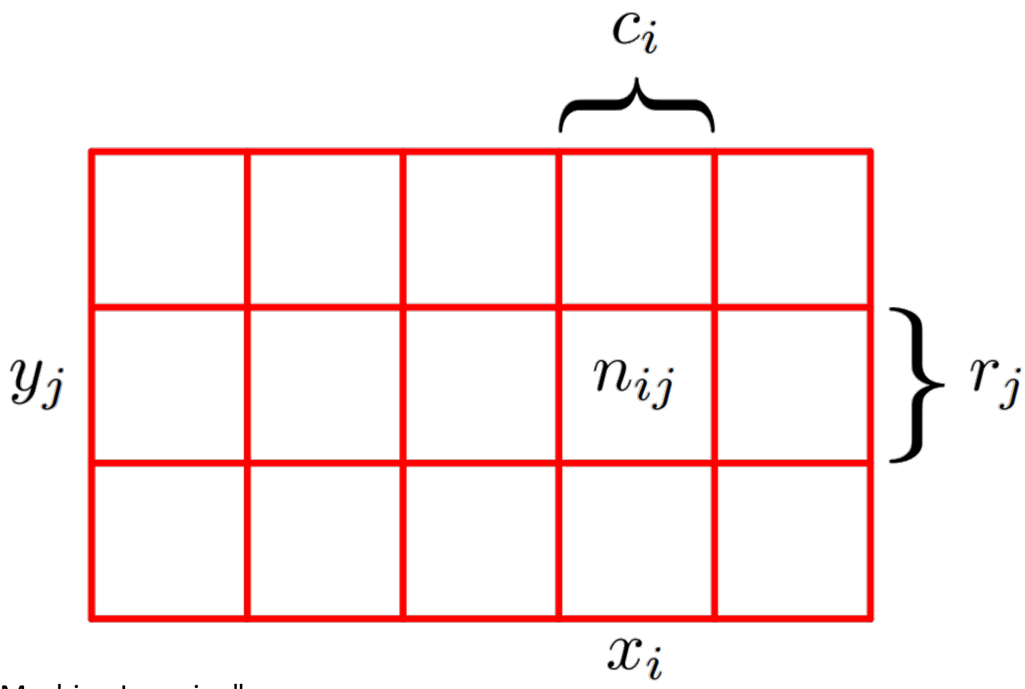
- To answer such questions, we need the elementary rules of probability:
  - the **sum rule**
  - the **product rule**

➢ Taken from Christopher M. Bishop, "Pattern Recognition and Machine Learning"

Joint Probability Distributions

- Consider we have two random variables X and Y
  - X can take any of the values x_i where $i = 1, \dots, M$
  - Y can take the values y_j where $j = 1, \dots, L$

- Consider a total of N samples
  - The number of such trials in which $X = x_i$ and $Y = y_j$ be n_{ij}
  - The number of trials in which X takes the value x_i is c_i
  - The number of trials in which Y takes the value y_j is r_j



➢ Taken from Christopher M. Bishop, "Pattern Recognition and Machine Learning"

Joint Probability Distributions

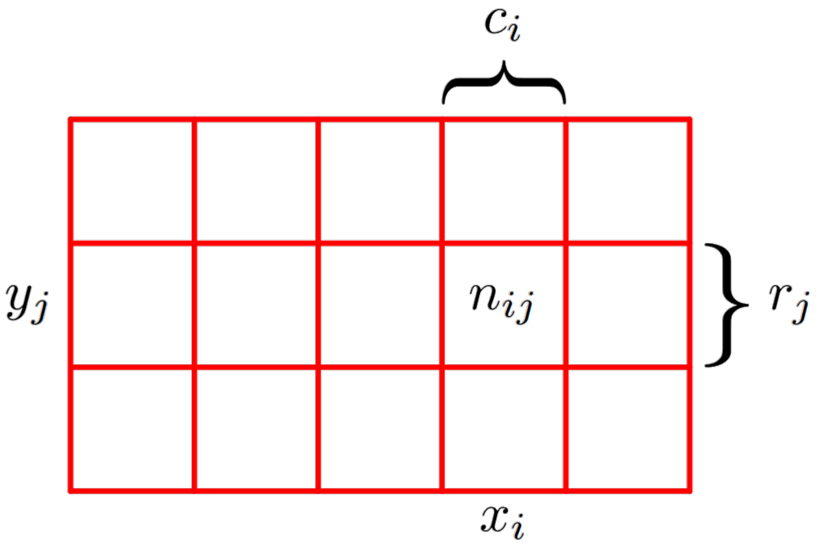
- The probability that $X = x_i$ and $Y = y_i$ is written as:

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

- This is called the joint probability distribution

- It is given by the fractions of point in cell i, j

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



➢ Taken from Christopher M. Bishop, "Pattern Recognition and Machine Learning"

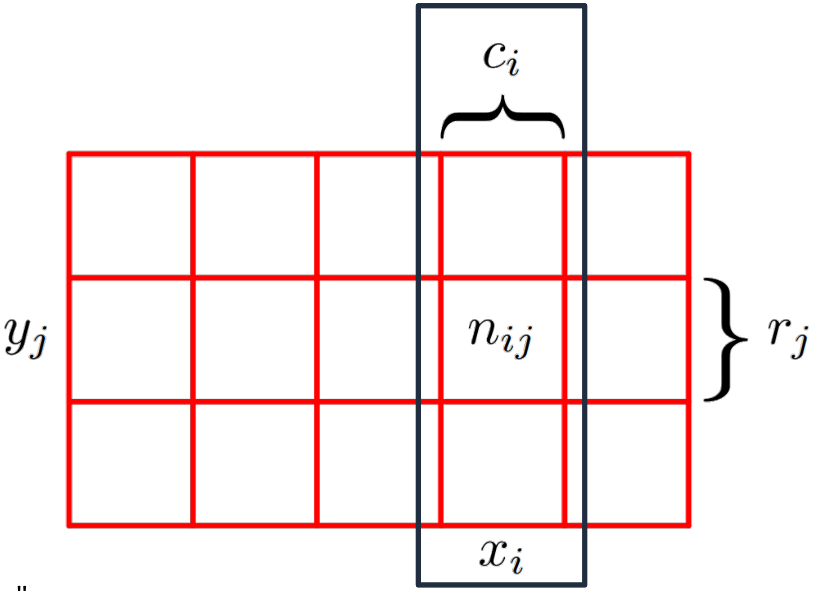
Sum Rule

- The probability that X takes the value x_i irrespective of the value of Y is written as $p(X = x_i)$ and is given by the fraction of points in column i

$$p(X = x_i) = \frac{c_i}{N}$$

- Since c_i is just the sum of the cells of the column, we can also write:

$$p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$



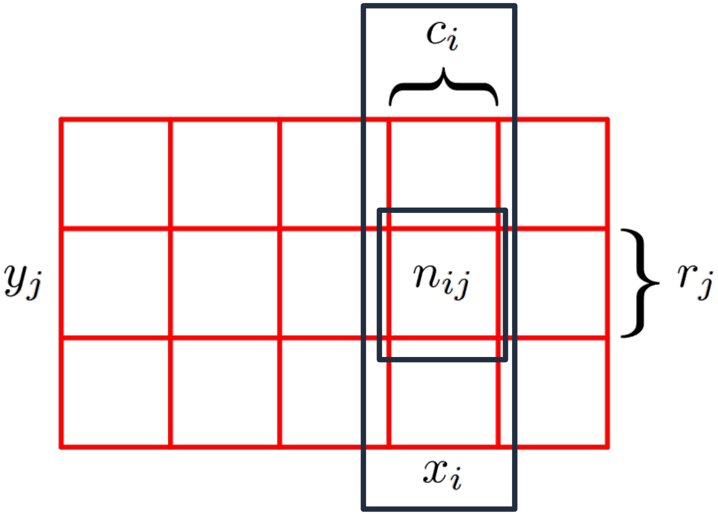
Note that $p(X = x_i)$ is sometimes called the marginal probability, because it is obtained by marginalizing, or summing out, the other variables (in this case Y).

➢ Taken from Christopher M. Bishop, "Pattern Recognition and Machine Learning"

Product Rule or Conditional Probabilities

- Now, let's only consider those cases, where $X = x_i$.
- We are now interested in the probability that $Y = y_j$ if X is already fixed to x_i
  (For example: What is the probability of sampling an apple, when we have selected the red box)
- We write this as:

$$p(Y = y_j | X_i = x_i) = \frac{n_{ij}}{c_i}$$



- With the results from before, we get:

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

➢ Taken from Christopher M. Bishop, "Pattern Recognition and Machine Learning"

Rules of Probability

- **Sum Rule**

$$p(X) = \sum_{Y} p(X, Y)$$

- **Product Rule**

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

(Note the more compact notation: We simply write $p(B)$ to denote a distribution over the random variable B, or $p(r)$ to denote the distribution evaluated for the particular value r, provided that the interpretation is clear from the context.)

➢ Taken from Christopher M. Bishop, "Pattern Recognition and Machine Learning"

Bayes theorem

- From the rules, we can directly derive Bayes theorem

$$\overbrace{P(Y|X)}^{\text{posterior}} = \frac{\overbrace{P(X|Y)}^{\text{likelihood}}\ \overbrace{P(Y)}^{\text{prior}}}{\underbrace{P(X)}_{\text{evidence}}}$$

- **Prior**: Our assumptions about Y before observing the data.
- **Likelihood**: The effect of the observed data X.
- **Posterior**: The uncertainty in Y after we have observed X.

"A Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule."

Bayes theorem in Action

- Consider a SPAM filter which classifies mail into:
  - Good: H (Ham)
  - Bad: S (Spam)
- We observe a message containing the word W "replica"
- We use Bayes theorem

$$p(S|W) = \frac{p(W|S)p(S)}{p(W|S)p(S) + p(W|H)p(H)} = \frac{p(W|S)p(S)}{p(W)}$$

  - $p(S|W)$ the probability that this message is SPAM
  - $p(S)$ the probability that any given message is SPAM (our **prior**)
  - $p(W|S)$ the probability that "replica" appears in SPAM (our **likelihood**)
  - $p(W)$ the probability that this word appears in any message (our **evidence**)

How does it work?

- We have trained the classifier, i.e., we have scanned through emails we know are SPAM or HAM and have calculated the following this:
    - 80% of the mail are SPAM, i.e. our prior $p(s) = 0.8$
        - That means, without any evidence, we are 80% certain, a message is SPAM
    - The probabilities of words occurring in SPAM and HAM

- Here, our filter again, in all its glory:

$$p(S|W) = \frac{p(W|S)p(S)}{p(W|S)p(S) + p(W|H)p(H)}$$

How does it work?

- We observe a completely irrelevant word: "the"
  - It appears in every singe message
  - $p(W|S) = 1$
  - $p(W|H) = 1$

- When we now apply the filter, we get:

$$p(S|W) = \frac{p(W|S)p(S)}{p(W|S)p(S) + p(W|H)p(H)}$$
$$= \frac{1 \cdot 0.8}{1 \cdot 0.8 + 1 \cdot 0.2} = 0.8$$

- That means, this useless evidence has neither strengthen nor weakened our prior believe

How does it work?

- Now we observe a typical SPAM word: "replica"
  - It appears in every fourth SPAM message, but only in every 100th HAM:
  - $p(W|S) = 0.25$
  - $p(W|H) = 0.01$

- When we now apply the filter, we get:

$$p(S|W) = \frac{p(W|S)p(S)}{p(W|S)p(S) + p(W|H)p(H)}$$

$$= \frac{0.25 \cdot 0.8}{0.25 \cdot 0.8 + 0.01 \cdot 0.2} = 0.99$$

- That means, the observed evidence, lead our posterior to be stronger than the prior!

How does it work?

- Now we observe a HAM, my name is spelled correct: "Röttger"
  - It appears as follows:
  - $p(W|S) = 0.05$
  - $p(W|H) = 0.75$

- When we now apply the filter, we get:

$$p(S|W) = \frac{p(W|S)p(S)}{p(W|S)p(S) + p(W|H)p(H)}$$

$$= \frac{0.05 \cdot 0.8}{0.05 \cdot 0.8 + 0.75 \cdot 0.2} = 0.21$$

- That means, the observed evidence, lead our posterior to less leaning towards SPAM!

Statistics

- Why Probability?
- Introduction to Probability
- Random Variables & Distributions
- Important Distribution Functions

Random Variables

- A **random variable** X is a variable that can take on different values randomly

- On its own, a random variable is just a description of the states that are possible;

- It must be coupled with a probability distribution that specifies how likely each of these states are.

- Random variables may be **discrete** or **continuous**

Probability Distributions

- A probability distribution is a description of how likely a random variable or a set of random variables is to take each of its possible states

- The way to describe the distribution depends on whether it is discrete or continuous

Probability Mass Functions

- The probability distribution over discrete variables is given by a probability mass function
- PMFs of variables are denoted by P and inferred from their argument, e.g., $P(x)$, $P(y)$
- They can act on many variables and is known as a joint distribution, written as $P(x, y)$

- To be a PMF it must satisfy:
  - Domain of P is the set of all possible states of x
  - $\forall x \in X, 0 \leq P(x) \leq 1$
  - $\sum_{x \in X} P(x) = 1$ (normalization)

Continuous Variables and PDFs

- When working with continuous variables, we describe probability distributions using probability density functions

- To be a pdf p must satisfy:
  - The domain of p must be the set of all possible states of X
  - $\forall x \in X, p(x) \geq 0$. Note, there is no requirement for $p(x) \leq 1$.
  - $\int p(x)dx = 1$

Expectation

- Expectation or expected value of $f(x)$ w.r.t. $P(X)$ is the average or mean value that f takes on when x is drawn from P

- For discrete variables:

$$E[X] = \sum_{x \in X} P(x) \cdot x$$

- For continuous variables:

$$E[X] = \int_x p(x) x \, dx$$

Variance

$$Var(X) = E[(x - E[X])^2]$$

- Measure how much the value of $f(x)$ vary from the expectation
- Low variance means values cluster around its expectations
- Square root of the variance is the standard deviation

Statistics

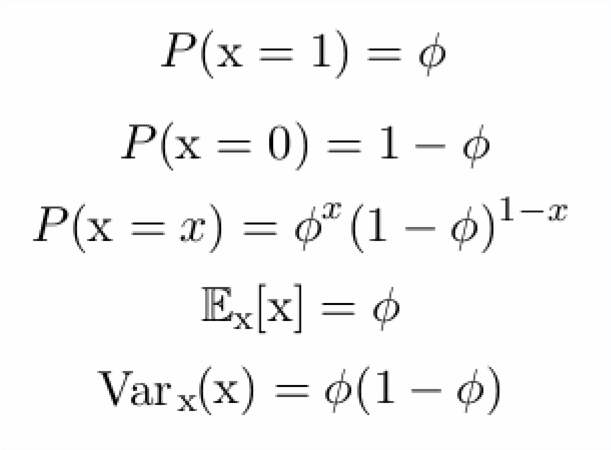
- Why Probability?
- Introduction to Probability
- Random Variables & Distributions
- Important Distribution Functions

Bernoulli Distribution

- Distribution over a single binary random variable

- It is controlled by a single parameter $\phi \in [0,1]$
  - Which gives the probability a random variable being equal to 1

- It has the following properties



$$P(\mathrm{x} = 1) = \phi$$

$$P(\mathrm{x} = 0) = 1 - \phi$$

$$P(\mathrm{x} = x) = \phi^{x}(1 - \phi)^{1-x}$$

$$\mathbb{E}_{\mathrm{x}}[\mathrm{x}] = \phi$$

$$\mathrm{Var}_{\mathrm{x}}(\mathrm{x}) = \phi(1 - \phi)$$

Multinoulli Distribution

- Distribution over a single discrete variable with k different states

- It is parameterized by a vector $\boldsymbol{p} \in [0,1]^{k-1}$
  - where p_i is the probability of the ith state
  - The final kth state’s probability is implicitly given by $1 - \mathbf{1}^T\boldsymbol{p}$
  - We must constrain $\mathbf{1}^T\boldsymbol{p} \leq 1$

- Multinoullis refer to distributions over categories
  - So we don’t assume state 1 has value 1, etc.
  - For this reason we do not usually need to compute the expectation or variance of multinoulli variables since the states are not necessarily ordered

Gaussian Distribution

- Probably one of the most commonly used distributions

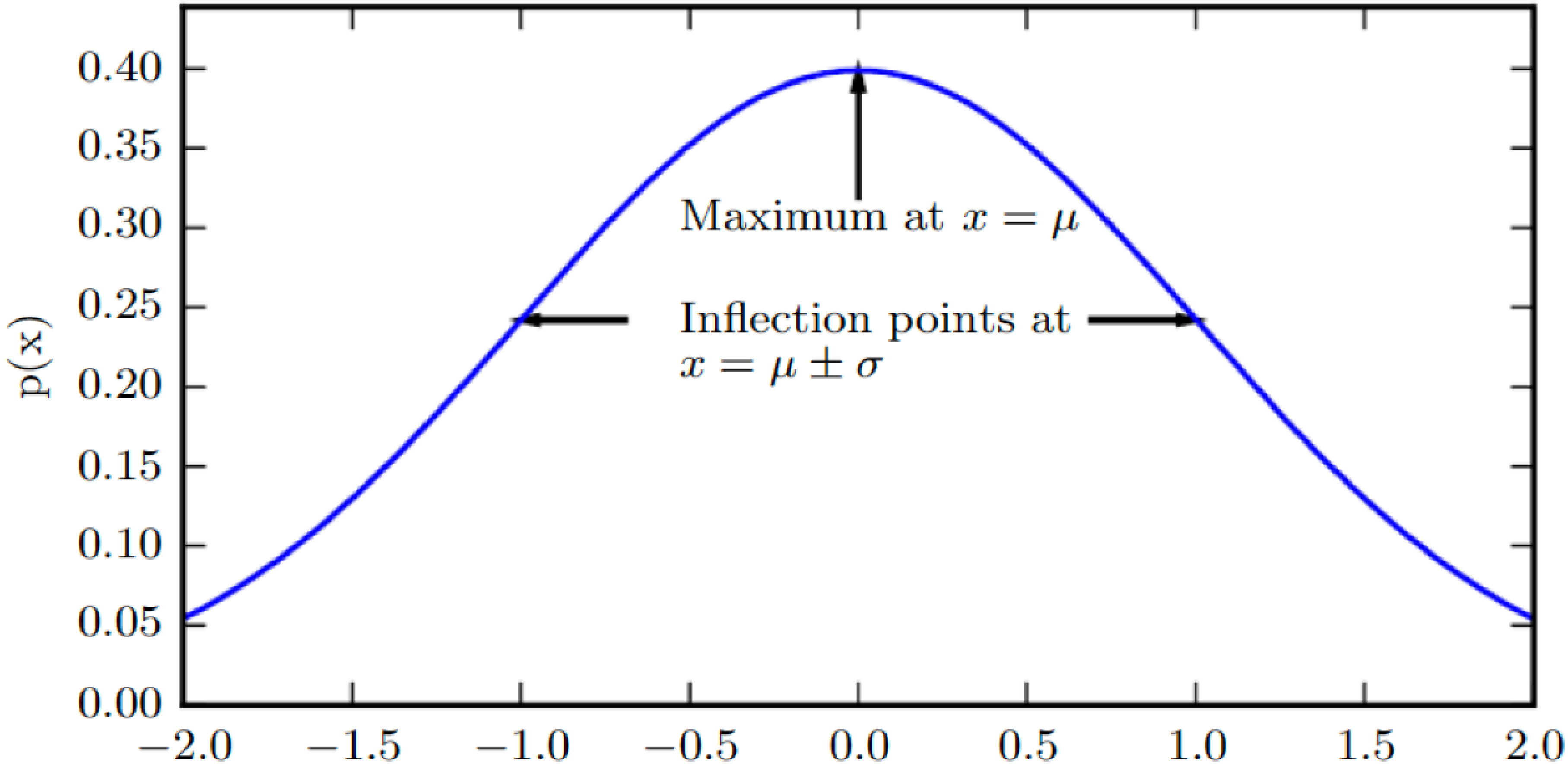
$$N(x;\mu,\sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

- Two parameters
  - μ gives the location of the central peak, which is also the mean of the distribution
  - The standard deviation is given by σ and variance by σ^2

- In case, this is evaluated frequently, sometimes parameterized with the inverse variance (or precision) β:

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

Gaussian Distribution, $\mu = 0, \sigma = 1$



Justifications for Normal Assumption

- 1. Central Limit Theorem
  - Many distributions we wish to model are truly normal
  - Sum of many independent distributions is normal
  - Can model complicated systems as normal even if components have more structured behavior

- 2. Maximum Entropy
  - Of all possible probability distributions with the same variance, normal distribution encodes the maximum amount of uncertainty over real numbers
  - Thus the normal distributions inserts the least amount of prior knowledge into a model

Multidimensional Gaussian Distributions

- The Gaussian Distribution can easily be extended to the multivariate case.
- Now, $\boldsymbol{x}$ and $\boldsymbol{\mu}$ are a vector, $\boldsymbol{\Sigma}$ a positive semidefinite symmetric matrix (the covariance matrix):

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

- Analogously, with the precision Matrix

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