Probability and Information Theory

Sargur N. Srihari
srihari@cedar.buffalo.edu

Topics in Probability and Information Theory

- Overview

1. Why Probability?
2. Random Variables
3. Probability Distributions
4. Marginal Probability
5. Conditional Probability
6. The Chain Rule of Conditional Probabilities
7. Independence and Conditional Independence
8. Expectation, Variance and Covariance
9. Common Probability Distributions
10. Useful Properties of Common Functions
11. Bayes Rule
12. Technical Details of Continuous Variables
13. Information Theory
14. Structured Probabilistic Models

Probability Theory and Information Theory

- Probability Theory
  - A mathematical framework for representing uncertain statements
  - Provides a means of quantifying uncertainty and axioms for deriving new uncertain statements
- Use of probability theory in artificial intelligence
  1. Tells us how AI systems should reason
     - So we design algorithms to compute or approximate various expressions using probability theory
  2. Theoretically analyze behavior of AI systems

Why Probability?

- Much of CS deals with entities that are certain
  - CPU executes flawlessly
    - Errors do occur but design need not be concerned
  - 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

- Need ability to reason with uncertainty
  - Beyond math statements true by definition, hardly any propositions are guaranteed
- Three 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 all variables are unobserved
  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

Can probability theory provide tools?

- 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

Logic and Probability

- Reasoning about uncertainty behaves the same way as frequentist probabilities
- Probability is an extension of logic to deal with uncertainty
- Logic provides rules for determining what propositions are implied to be true or false
- Probability theory provides rules for determining the likelihood of a proposition being true given the likelihood of other propositions

Random Variables

- Variable that can take different values randomly
- Scalar random variable denoted x
- Vector random variable is denoted in bold as $\mathbf{x}$
- Values of r.v.s denoted in italics x or $\boldsymbol{x}$
  - Values denoted as $\mathrm{Val}(\mathrm{x})=\{x_1,x_2\}$
- Random variable must has a probability distribution to specify how likely the states are
- Random variables can be discrete or continuous
  - Discrete values need not be integers, can be named states
  - Continuous random variable is associated with a real value

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

Discrete Variables and PMFs

- 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., $P(\mathrm{x})$, $P(\mathrm{y})$
- 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:
  1. Domain of P is the set of all possible states of x
  2. $\forall x \in \mathrm{x}, 0 \leq P(x) \leq 1$ It is not necessary for $P(x) \leq 1$
  3. Normalization $\sum_{x \in \mathrm{x}} P(x) = 1$

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.
  - $\forall x \in \mathrm{x}, p(x) \geq 0$. Note that we do not require $p(x) \leq 1$.
  - $\int p(x)dx = 1$.

Marginal Probability

- Sometimes we know the joint distribution of several variables
- And we want to know the distribution over some of them
- It can be computed using

$$\forall x \in \mathrm{x}, P(\mathrm{x} = x) = \sum_{y} P(\mathrm{x} = x, \mathrm{y} = y)$$

$$p(x) = \int p(x, y) dy$$

Conditional Probability

- We are often interested in the probability of an event given that some other event has happened
- This is called conditional probability
- It can be computed using

$$P(\mathrm{y} = y \mid \mathrm{x} = x) = \frac{P(\mathrm{y} = y, \mathrm{x} = x)}{P(\mathrm{x} = x)}$$

Chain Rule of Conditional Probability

- Any probability distribution over many variables can be decomposed into conditional distributions over only one variable

$$P(\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(n)}) = P(\mathrm{x}^{(1)}) \Pi_{i=2}^{n} P(\mathrm{x}^{(i)} \mid \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(i-1)})$$

- An example with three variables

$$\begin{aligned} P(\mathrm{a}, \mathrm{b}, \mathrm{c}) &= P(\mathrm{a} \mid \mathrm{b}, \mathrm{c}) P(\mathrm{b}, \mathrm{c}) \\ P(\mathrm{b}, \mathrm{c}) &= P(\mathrm{b} \mid \mathrm{c}) P(\mathrm{c}) \\ P(\mathrm{a}, \mathrm{b}, \mathrm{c}) &= P(\mathrm{a} \mid \mathrm{b}, \mathrm{c}) P(\mathrm{b} \mid \mathrm{c}) P(\mathrm{c}) \end{aligned}$$

Independence & Conditional Independence

- Independence: $\boxed{\mathrm{x} \perp \mathrm{y}}$
  - Two variables x and y are independent if their probability distribution can be expressed as a product of two factors, one involving only x and the other involving only y

$$\forall x \in \mathrm{x}, y \in \mathrm{y},\ p(\mathrm{x} = x, \mathrm{y} = y) = p(\mathrm{x} = x)p(\mathrm{y} = y)$$

- Conditional Independence: $\boxed{\mathrm{x} \perp \mathrm{y} \mid \mathrm{z}}$
  - Two variables x and y are independent given variable z, if the conditional probability distribution over x and y factorizes in this way for every z

$$\forall x \in \mathrm{x}, y \in \mathrm{y}, z \in \mathrm{z},\ p(\mathrm{x} = x, \mathrm{y} = y \mid \mathrm{z} = z) = p(\mathrm{x} = x \mid \mathrm{z} = z)p(\mathrm{y} = y \mid \mathrm{z} = z)$$

Expectation

- Expectation or expected value of $f(x)$ wrt $P(\mathrm{x})$ is the average or mean value that f takes on when x is drawn from P
- For discrete variables

$$\mathbb{E}_{\mathbf{x}\sim P}[f(x)] = \sum_{x} P(x) f(x)$$

- For continuous variables

$$\mathbb{E}_{\mathbf{x}\sim p}[f(x)] = \int p(x) f(x) dx.$$

Variance

- Variance gives a measure of how much the values of a function of a random variable x vary as we sample x from a probability distribution

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

- When the variance is low, values of $f(x)$ cluster around its expected value
- The square root of the variance is known as the standard deviation

Covariance

- Covariance measures how two values are linearly related, as well as scale of variables

$$\mathrm{Cov}(f(x), g(y)) = \mathbb{E}\left[(f(x) - \mathbb{E}\left[f(x)\right])\left(g(y) - \mathbb{E}\left[g(y)\right]\right)\right]$$

  - High absolute values of covariance:
    - Values change very much & are both far from their mean
  - If sign is positive
    - Both variables take relatively high values far from mean
  - If sign is negative
    - One var. takes on high values & another takes low values
- Correlation normalizes each variable
  - Measures only how variables are related
    - Not affected by scale of variables

Independence stronger than covariance

- Covariance & independence are related but not same
- Zero covariance is necessary for independence
  - Independent variables have zero covariance
  - Variables with non-zero covariance are dependent
- Independence is a stronger requirement
  - They not only must *not* have linear relationship (zero covariance)
  - They must not have nonlinear relationship either

Ex: Dependence with zero covariance

- Suppose we sample real number x from $\mathrm{U}[-1,1]$
- Next sample a random variable s
  - with prob ½ we choose $s=1$ otherwise $s = -1$
- Generate random variable y assigning $y = sx$
  - i.e., $y=-x$ or $y=x$ depending on s
  - Clearly x and y are not independent
    - Because x completely determines magnitude of y
- However $\mathrm{Cov}(x,y)=0$
  - Because when x has a high value y can be high or low depending on s

Common Probability Distributions

- Several simple probability distributions are useful in may contexts in machine learning
  - Bernoulli over a single binary random variable
  - Multinoulli distribution over a variable with k states
  - Gaussian distribution
  - Mixture distribution

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 with k finite
- It is parameterized by a vector $\boldsymbol{p} \in [0,1]^{k-1}$
  - where p_i is the probability of the i^{th} state
  - The final k^{th} state's probability is given by $1 - \mathbf{1}^\top \boldsymbol{p}$.
  - We must constrain $\mathbf{1}^\top \boldsymbol{p} \leq 1$
- 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

Gaussian Distribution

- Most commonly used distribution over real numbers is the Gaussian or normal distribution

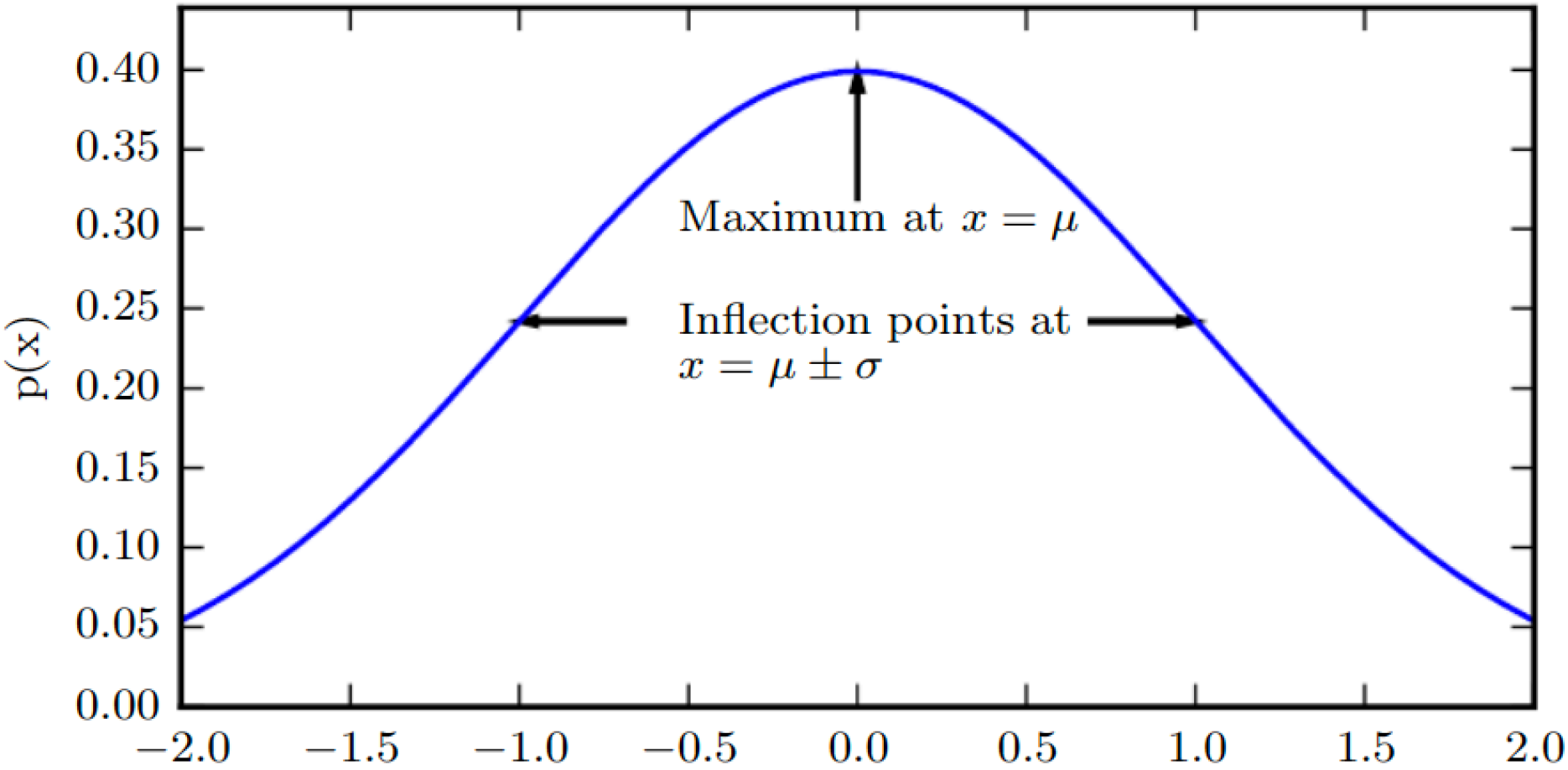
$$\mathcal{N}(x;\mu,\sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

- The two parameters $\mu \in \mathbb{R}$ and $\sigma \in (0,\infty)$
  - Control the normal distribution
    - Parameter μ gives the coordinate of the central peak
    - This is also the mean of the distribution $\mathbb{E}[x] = \mu$.
    - The standard deviation is given by σ and variance by σ^2
- To evaluate PDF need to square and invert σ.
    - To evaluate PDF often, more efficient to use precision or inverse variance

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

Standard normal 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 nos.
   - Thus the normal distributions inserts the least amount of prior knowledge into a model

Normal distribution in R^{n}

- A multivariate normal may be parameterized with a positive definite symmetric matrix Σ

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

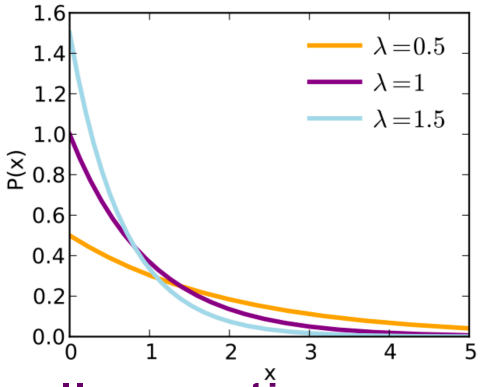
  - μ is a vector-valued mean, Σ is the covariance matrix
- If we wish to evaluate the pdf for many different values of parameters, inefficient to invert Σ to evaluate the pdf. Instead use precision matrix β

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

Exponential and Laplace Distributions

- In deep learning we often want a distribution with a sharp peak at $x=0$.
  - Accomplished by *exponential*

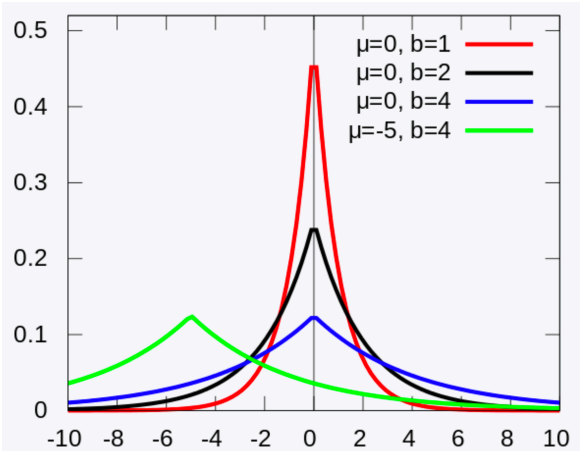
$$p(x;\lambda) = \lambda \mathbf{1}_{x\geq 0} \exp(-\lambda x)$$



  - Indicator $\mathbf{1}_{x\geq 0}$ assigns probability zero to all negative x

- *Laplace* distribution is closely-related
  - It allows us to place a sharp peak at arbitrary μ

$$\text{Laplace}(x;\mu,\gamma) = \frac{1}{2\gamma}\exp\left(-\frac{|x-\mu|}{\gamma}\right)$$



Dirac Distribution

- To specify that mass clusters around a single point, define pdf using Dirac delta function $\delta(x)$:

$$p(x) = \delta(x - \mu)$$

  - Dirac delta: zero everywhere except 0, yet integrates to 1
  - It is not an ordinary function. Called a *generalized function* defined in terms of properties when integrated

- By defining $p(x)$ to be δ shifted by $-\mu$ we obtain an infinitely narrow and infinitely high peak of probability mass where $x = \mu$
- Common use of Dirac delta distribution is as a component of an empirical distribution

Empirical Distribution

- Dirac delta distribution is used to define an empirical distribution over continuous variables

$$\hat{p}(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m}\delta(\boldsymbol{x}-\boldsymbol{x}^{(i)})$$

  - which puts probability mass $1/m$ on each of m points $\boldsymbol{x}^{(1)},..\boldsymbol{x}^{(m)}$ forming a given dataset
- For discrete variables, the situation is simpler
  - Probability associated with each input value is the empirical frequency of that value in the training set
- Empirical distribution is the probability density that maximizes the likelihood of training data

Mixtures of Distributions

- A mixture distribution is made up of several component distributions
- On each trial, the choice of which component distribution generates the sample is determined by sampling a component identity from a multinoulli distribution:

$$P(\mathrm{x}) = \sum_i P(\mathrm{c} = i) P(\mathrm{x} \mid \mathrm{c} = i)$$

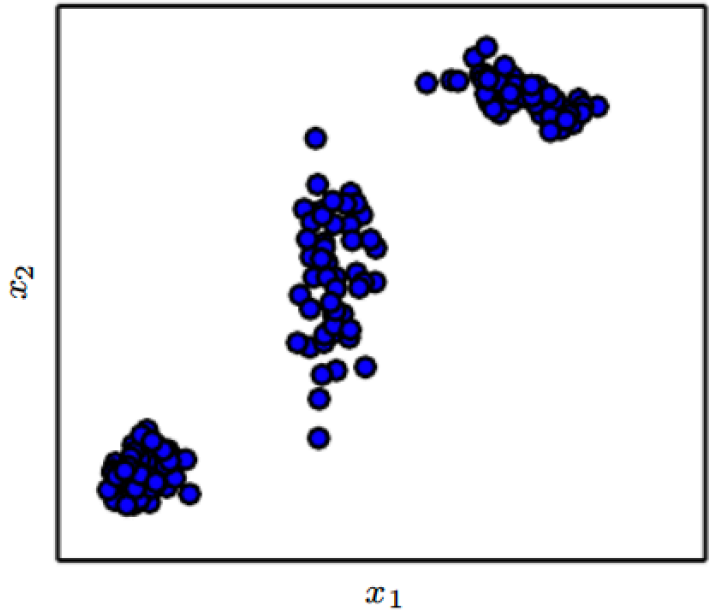
  - where $P(\mathrm{c})$ is a multinoulli distribution
- Ex: empirical distribution over real-valued variables is a mixture distribution with one Dirac component for each training example

Creating richer distributions

- Mixture model is a strategy for combining distributions to create a richer distribution
  - PGMs allow for more complex distributions
- Mixture model has concept of a latent variable
  - A latent variable is a random variable that we cannot observe directly
    - Component identity variable c of the mixture model provides an example
    - Latent vars relate to x through joint $P(\mathrm{x},\mathrm{c})=P(\mathrm{x}|\mathrm{c})P(\mathrm{c})$
      - $P(\mathrm{c})$ is over latent variables and
      - $P(\mathrm{x}|\mathrm{c})$ relates latent variables to the visible variables
      - Determines shape of the distribution $P(\mathrm{x})$ even though it is possible to describe $P(\mathrm{x})$ without reference to latent variable

Gaussian Mixture Models

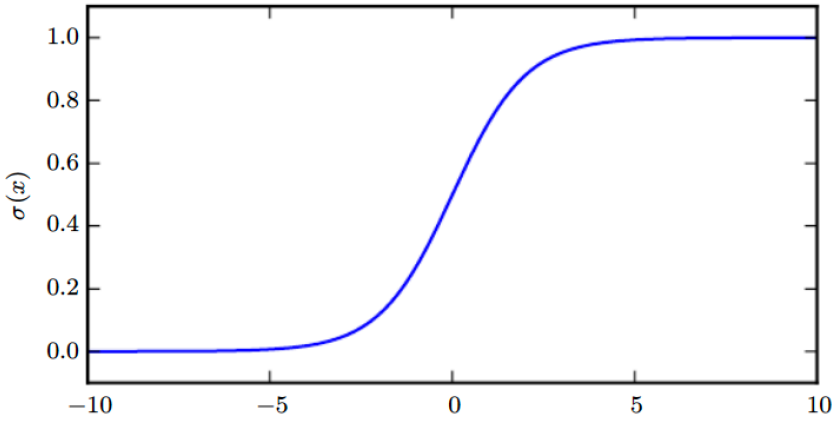
- Components $p(\mathrm{x}|\mathrm{c}=i)$ are Gaussian
- Each component has a separately parameterized mean $\boldsymbol{\mu}^{(i)}$ and covariance $\boldsymbol{\Sigma}^{(i)}$
- Any smooth density can be approximated with enough components
- Samples from a GMM:
  - 3 components
    - Left: isotropic covariance
    - Middle: diagonal covariance
      - Each component controlled
    - Right: full-rank covariance



Useful properties of common functions

- Certain functions arise with probability distributions used in deep learning
- Logistic sigmoid $\sigma(x) = \frac{1}{1+\exp(-x)}$
  - Commonly used to produce the ϕ parameter of a Bernoulli distribution because its range is $(0,1)$

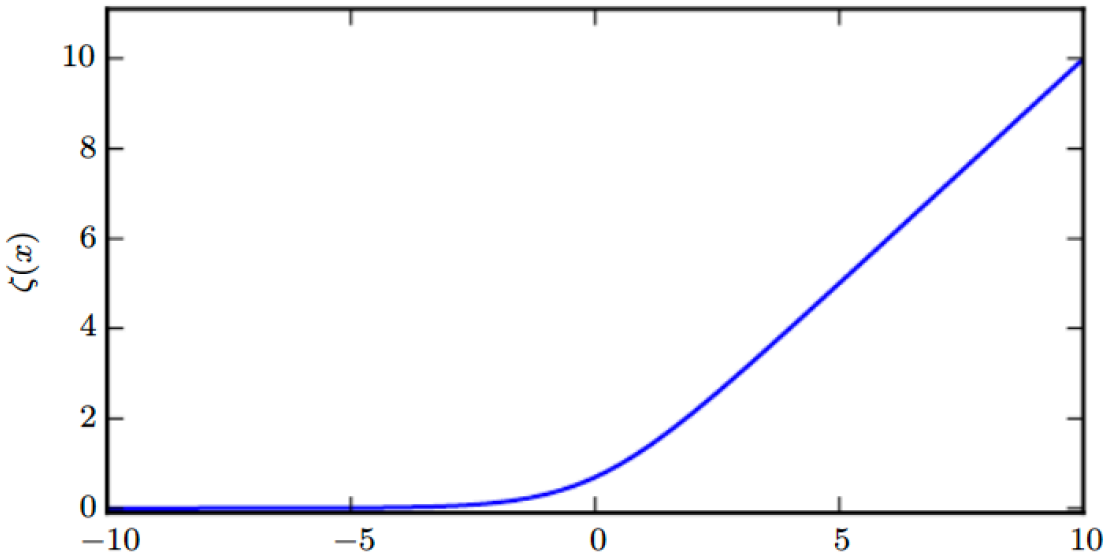
$$P(\mathrm{x}=1) = \phi$$
$$P(\mathrm{x}=0) = 1-\phi$$
$$P(\mathrm{x}=x) = \phi^{x}(1-\phi)^{1-x}$$
$$\mathbb{E}_{\mathrm{x}}[\mathrm{x}] = \phi$$
$$\mathrm{Var}_{\mathrm{x}}(\mathrm{x}) = \phi(1-\phi)$$



  - It saturates when x is very small/large
    - Thus it is insensitive to small changes in input

Softplus Function

- It is defined as $\zeta(x) = \log(1 + \exp(x))$
  - Softplus is useful for producing the β or σ parameter of a normal distribution because its range is $(0,\infty)$
  - Also arises in manipulating sigmoid expressions
- Name arises as smoothed version of $x^+ = \max(0,x)$



Useful identities

$$\sigma(x) = \frac{\exp(x)}{\exp(x) + \exp(0)}$$

$$\frac{d}{dx}\sigma(x) = \sigma(x)(1 - \sigma(x))$$

$$1 - \sigma(x) = \sigma(-x)$$

$$\log \sigma(x) = -\zeta(-x)$$

$$\frac{d}{dx}\zeta(x) = \sigma(x)$$

$$\forall x \in (0, 1),\ \sigma^{-1}(x) = \log\left(\frac{x}{1 - x}\right)$$

$$\forall x > 0,\ \zeta^{-1}(x) = \log\left(\exp(x) - 1\right)$$

$$\zeta(x) = \int_{-\infty}^{x} \sigma(y)dy$$

$$\zeta(x) - \zeta(-x) = x$$

Bayes' Rule

- We often know $P(\mathrm{y}|\mathrm{x})$ and need to find $P(\mathrm{x}|\mathrm{y})$
  - Ex: in classification, we know $P(x|C_i)$ and need to find $P(C_i|x)$
- If we know $P(\mathrm{x})$ then we can get the answer as

$$P(\mathrm{x} \mid \mathrm{y}) = \frac{P(\mathrm{x})P(\mathrm{y} \mid \mathrm{x})}{P(\mathrm{y})}$$

  - Although $P(\mathrm{y})$ appears in formula, it can be computed as

$$P(\mathrm{y}) = \sum_{x} P(\mathrm{y} \mid x)P(x)$$

    - Thus we don't need to know $P(\mathrm{y})$
- Bayes' rule is easily derived from the definition of conditional probability