SDU Summer School

Deep Learning

Summer 2018

Probability and Information Theory

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

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

- 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 $p(x) \leq 1$.
  - $\int p(x)dx = 1$

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:
  - In the discrete case:

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

  - In the continues case:

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

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

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

- Bayes Theorem

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

Bayes Theorem

- The Disease:
  - Let's assume we invented a test to check if an individual has a disease D
  - The disease is rare, we have 100 sick in 1, 000, 000 people
- Our Test:
  - When presented with a sick patient, we receive a positive result in 99% of the cases
  - When presented with a healthy person, we receive a (falsely) positive result in 1% of the cases
- Question:
  - We are tested, and shockingly the result is positive!
  - How likely is it, that we are actually sick?

Bayes Theorem

- $\Pr[S] = \frac{100}{1000000} = 0.0001$ is the probability a random person is sick
- $\Pr[P] = \Pr[P|S] \cdot \Pr[S] + \Pr[P|\overline{S}] \Pr[\overline{S}] =$
  $0.99 \cdot 0.0001 + 0.01 \cdot 0.9999 = 0.0101$
  - The Probability that we receive a positive test
  - Is the probability of a true positive + the probability of a false negative

- We are interested in $\Pr[S|P]$

Bayes Theorem

- $\Pr[S] = \frac{100}{1000000} = 0.0001$ is the probability a random person is sick
- $\Pr[P] = \Pr[P|S] \cdot \Pr[S] + \Pr[P|\overline{S}] \Pr[\overline{S}] = 0.99 \cdot 0.0001 + 0.01 \cdot 0.9999 = 0.0101$
  - The Probability that we receive a positive test
  - Is the probability of a true positive + the probability of a false negative

- We are interested in $\Pr[S|P]$
- Lets use Bayes Theorem:

$$\Pr[S|P] = \frac{\Pr[P|S] \Pr[S]}{\Pr[P]} = \frac{0.99 \cdot 0.0001}{0.0101} = 0.01$$

- Even with a positive result, the chance that we are actually sick is only 1%
- Compared to the 99% that a random test is correct

Conditional Probability & Independence

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

$$P\left(X^{(1)}, \ldots, X^{(n)}\right) = P\left(X^{(1)}\right) \prod_{i=2}^{n} P(X^{(i)} | X^{(1)}, \ldots, X^{(i-1)})$$

- For example:

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

- Two variables x and y are **independent** if their probability distribution can be expressed as a product of two factors:

$$\forall x \in X, y \in Y, p(X = x, Y = y) = p(X = x) \cdot p(Y = y)$$

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:

$$E[f(x)] = \sum_x P(x) f(x)$$

- For continuous variables:

$$E[f(x)] = \int p(x) f(x) dx$$

Variance

$$Var\big(f(x)\big) = E[(f(x) - E[f(x)])^2]$$

- 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

Covariance

- Covariance measures how two values are **linearly** related:

$$Cov\big(f(x), g(y)\big) = E[(f(x) - E[f(x)])(g(y) - E[g(y)])]$$

- Interpretation:
  - High absolute values of covariance: Values change very much & are both far from their mean
  - Positive: High values of $f(x)$ coincide with high values of $g(x)$
  - Negative: High values of $f(x)$ coincide with low values of $g(x)$

- Covariance & independence are related but not same
  - Zero covariance is necessary for independence but not sufficient
  - They must not have nonlinear relationship either

Independence vs. Covariance

- 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

Dependence with zero covariance

- Suppose we sample real number x from $U[-1,1]$
- Next sample a random variable s
  - with prob $^1/_2$ 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
  - x completely determines magnitude of y
- However $Cov(x, y) = 0$
  - Because when x has a high value y can be high or low depending on s

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

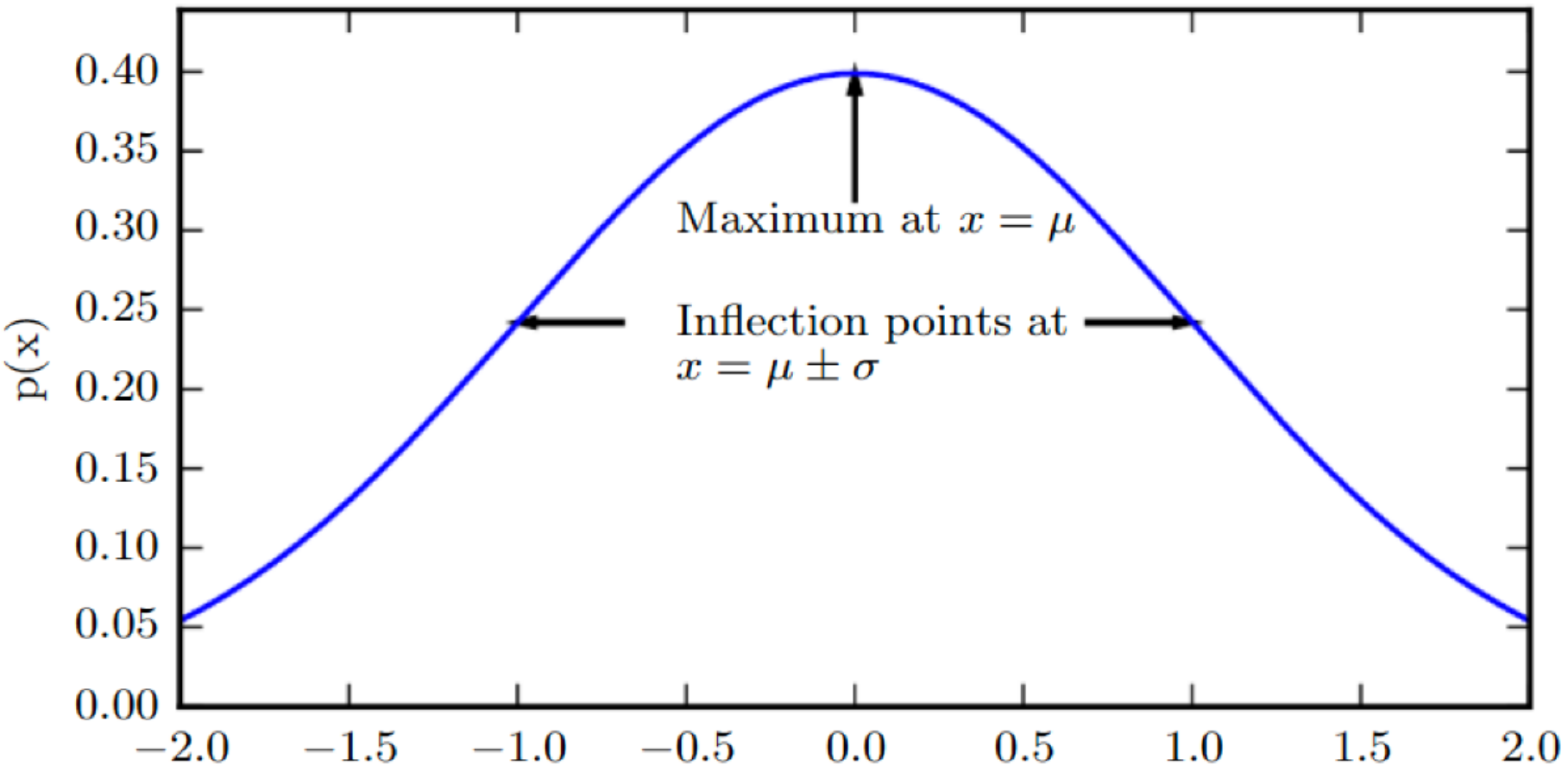
$$N(x;\mu,\sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

- 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

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