Probability, etc.

CS 4277 Deep Learning

Kennesaw State University

Probability[1]

Probability theory: quantification and manipulation of uncertainty.

- Epistemic, a.k.a. systematic uncertainty: we only see data sets of finite size
- Aleatoric, a.k.a. intrinsic, stochastic uncertainty: noise – we only observe partial information

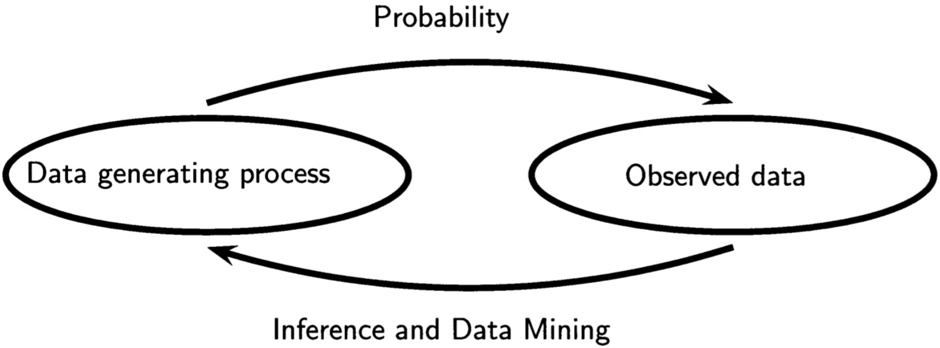
60%

40%

[1]Follows Chapter 2 of Deep Learning Foundations and Concepts

Probability in Machine Learning

We observe data generated by a random process.



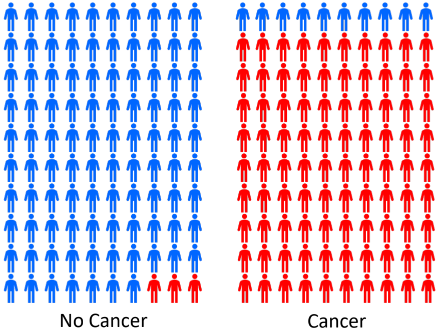
[2]

We make some assumptions about the data generating function and infer its parameters using samples from the process (training data).

[2] All of Statistics

A Medical Screening Example

A cancer with occurence rate of 1% (.01) has a "90% accurate" test, and:



False positive rate: .03, False negative rate: 0.10

Questions:

- If we screen someone, what is the probability that they test positive?
- If someone tests positive, what is the probability that they have cancer?

We'll return to these questions after we develop some analysis tools.

Joint Probability

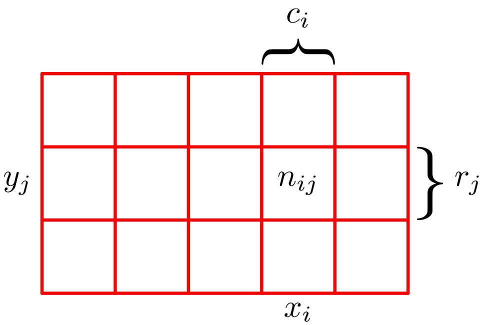
Let X and Y be *random* (a.k.a. *stochastic*) variables and

- $\{x_i\}_{i=1}^{L}$
- $\{y_j\}_{j=1}^{M}$
- N trials in which we sample X and Y
- n_{ij} is number of trials in which $X = x_i$ and $Y = y_j$
- c_i is the number of trials in which $X = x_i$, for all ys
- r_j is the number of trials in which $Y = y_j$, for all xs

Then the joint probability of observing x_i and y_j is

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

We can visualize this event with the grid diagram on the right. Note that we're always observing events where both random variables have values, e.g., when we screen a person for cancer we're observing a joint event of two random variables: the test result and the actual existence of cancer.



The Sum Rule

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

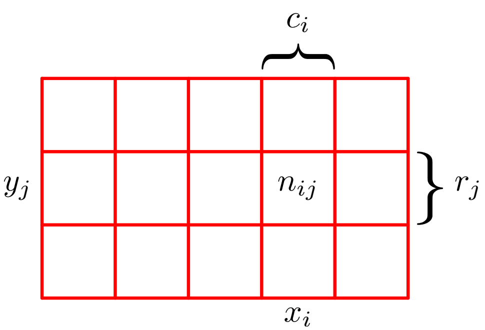
Notice that the number of instances in column i, c_i, is the sum of instances in each cell having n_{ij} instances, so $c_i = \sum_j n_{ij}$. Recalling that

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

we have

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

This is the *sum rule*, which is also called the marginal probability becuase we sum over the other variable and write the sum in the margin of the table.



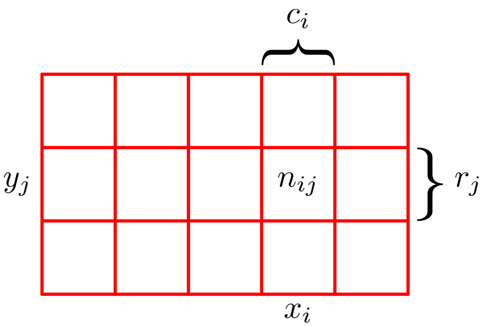
Conditional Probability

If we consider trials in which $X = x_i$, the fraction of those trials in which $Y = y_j$ is written

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

We call this the *conditional probability* of $Y = y_j$ given $X = x_i$, which is the fraction of points in column i that fall in cell i, j so:

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



The Product Rule

Given the previous definitions for conditional probabilities and marginal probabilities, we can derive a formula for joint probabilities:

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

This is the *product rule*.

We can summarize the sum and product rules with a more compact notation:

$$\text{Sum rule: } p(X) = \sum_Y p(X, Y)$$

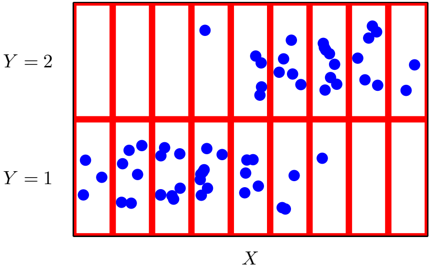
$$\text{Product rule: } p(X, Y) = p(Y|X)p(X)$$

These two rules underlie all the probabilistic machinery we'll use in this course.

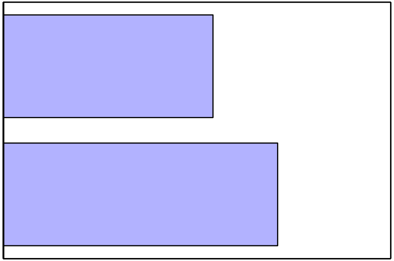
KENNESAW STATE UNIVERSITY

Visualizing Joint Distributions

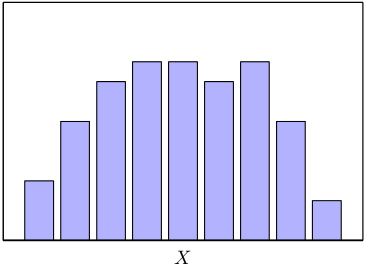
$p(X, Y)$



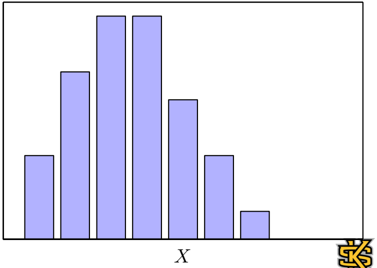
$p(Y)$



$p(X)$



$p(X|Y = 1)$



Bayes' Theorem

Using the symmetry $p(x, y) = p(Y, X)$ and the product rule:

$$\begin{aligned} p(X, Y) &= p(Y, X) \\ p(Y|X)p(X) &= p(X|Y)p(Y) \\ p(Y|X) &= \frac{p(X|Y)p(Y)}{p(X)} \end{aligned}$$

where the denominator $p(X)$ is a normalizing constant (what's that?):

$$p(X) = \sum p(X|Y)p(Y)$$

This is called *Bayes' Theorem* or *Bayes' Rule.*

We use Bayes' Theorem to update our beliefs after observing evidence. For example:

- Before we run the test, the *prior probability* that someone has cancer is $p(C)$
- After we run the test, we use Bayes' Theorem to calculate the *posterior probability* $p(C|T)$

The *posterior probability* is our new belief after a Bayesian update.

KENNESAW STATE UNIVERSITY

Analysis of Medical Screening Example

With our probabilistic machinery we can now analyze our cancer screening example. First, we model the problem in the language of Bayesian probability theory:

$$
\begin{aligned}
p(C=1) &= \frac{1}{100} && \text{(Prior probability that someone has cancer)} \\
p(C=0) &= \frac{99}{100} && \text{(Prior probability that someone has no cancer)} \\
p(T=1|C=1) &= \frac{90}{100} && \text{(Conditional probability of positive test given cancer)} \\
p(T=0|C=1) &= \frac{10}{100} && \text{(Conditional probability of negative test given cancer)} \\
p(T=1|C=0) &= \frac{3}{100} && \text{(Conditional probability of positive test given no cancer)} \\
p(T=0|C=0) &= \frac{97}{100} && \text{(Conditional probability of negative test given no cancer)}
\end{aligned}
$$

Now we can answer the two questions we posed at the outset:

- If we screen someone, what is the probability that they test positive?
- If someone tests positive, what is the probability that they have cancer?

KENNESAW STATE UNIVERSITY

Analysis of Medical Screening Example

$$
\begin{aligned}
p(C=1) &= \frac{1}{100} \\
p(C=0) &= \frac{99}{100} \\
p(T=1|C=1) &= \frac{90}{100} \\
p(T=0|C=1) &= \frac{10}{100} \\
p(T=1|C=0) &= \frac{3}{100} \\
p(T=0|C=0) &= \frac{97}{100}
\end{aligned}
$$

If we screen someone, probability that they test positive:

$$
\begin{aligned}
p(T=1) &= p(T=1|C=0)p(C=0) + p(T=1|C=1)p(C=1) \\
&= \frac{3}{100} \times \frac{99}{100} + \frac{90}{100} \times \frac{1}{100} \\
&= \frac{387}{10,000} \\
&= .0387
\end{aligned}
$$

If someone tests positive, probability they have cancer:

$$
\begin{aligned}
p(C=1|T=1) &= \frac{p(T=1|C=1)p(C=1)p(C=1)}{p(T=1)} \\
&= \frac{90}{100} \times \frac{1}{100} \times \frac{10,000}{387} \\
&= \frac{90}{387} \\
&\approx 0.23
\end{aligned}
$$

Independent Variables

If the joint distribution factorizes into the product of the marginals:

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

Then we say that X and Y are *independent*. So

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

and

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

Question: in our cancer screening example, is the probability of a positive test independent of whether a person has cancer?

KENNESAW STATE UNIVERSITY

Probability Densities

For continuous valus we need different probability rules, because the probability of any precise real number is effectively zero.

The probability density of a variable x falling in the interval $x + \delta x$ is $p(x)\delta x$ for $\delta x \to 0$. So the probability that x will be in the interval (a, b) is:

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

Just as a discrete probability is non-negative and a distribution must sum to 1, continuous probability densities must satisfy:

$$p(x) \geq 0$$

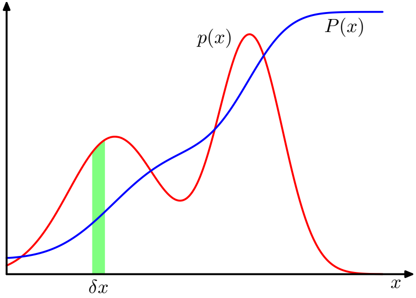
$$\int_{-\infty}^{\infty} p(x)\, dx = 1$$

KENNESAW STATE UNIVERSITY

Continuous CDF

The probability that x lies in the interval $(-\infty, z)$ is given by the cumulative distribution function (CDF):

$$P(z) = \int_{-\infty}^{z} p(x)\,dx.$$



The continuous CDF satisfies:

$$P'(x) = p(x)$$

KENNESAW STATE UNIVERSITY

Joint Probability Densities

Given $\boldsymbol{x} = (x_1, \cdots, x_D)$, the probability density $p(\boldsymbol{x}) = p(x_1, \cdots, x_D)$ where the probability of $\boldsymbol{x}$ falling in an infinitesimal volume $\delta\boldsymbol{x}$ is given by $p(\boldsymbol{x})\delta x$, so we have the multivariate density over the whole of $\boldsymbol{x}$ space is:

$$p(\boldsymbol{x}) \geq 0$$

$$\int p(\boldsymbol{x}) d\boldsymbol{x} = 1$$

KENNESAW STATE UNIVERSITY

Summary of Discrete and Continuous Probability Rules

Given discrete random variables X and Y:

$$\text{Sum rule: } p(X) = \sum_Y p(X, Y)$$

$$\text{Product rule: } p(X, Y) = p(Y|X)p(X)$$

Bayes' Theorem

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

with normalizing constant:

$$p(X) = \sum p(X|Y)p(Y)$$

Given continuous random variables $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$\text{Sum rule: } p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{y})d\boldsymbol{y}$$

$$\text{Product rule: } p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})$$

Bayes' Theorem:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})}{p(\boldsymbol{x})}$$

with normalizing constant:

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})d\boldsymbol{y}$$

KENNESAW STATE UNIVERSITY

Distributions

Uniform over a finite region (improper over infinite region because can't be normalized):

$$p(x) = \frac{1}{(d-c)}, x \in (c, d)$$

Exponential:

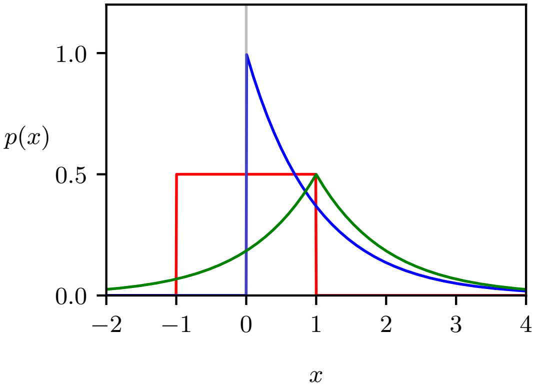
$$p(x|\lambda) = \lambda \exp(-\lambda x), x \geq 0$$

Laplace distribution creates a peak at μ:

$$p(x|\mu, \gamma) = \frac{1}{2\gamma} \exp(-\frac{|x-\mu|}{\gamma})$$

Dirac delta function is defined to be zero everywhere except at $x = \mu$, creating an infinitely tall spike at $x = \mu$:

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



- Red is uniform over $(-1, 1)$
- Blue is exponential with $\lambda = 1$
- Green is Laplace with $\mu = 1$ and $\gamma = 1$

Dirac Delta Function

The Dirac delta function

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

is interesting because if you have data ponits $\mathcal{D} = \{x_1, \cdots, x_N\}$ you can put a Dirac Delta function centered at each point to construct the *empirical distribution*:

$$p(x|\mathcal{D}) = \frac{1}{N} \sum_{n=1}^{N} \delta(x - x_n)$$

Expectations

The *expected value* or *mean* or *forst moment* of a random variable X is the weighted average of a function $f(x)$ under some probability distribution $p(x)$.

for discrete variables:

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

for coninuous variables:

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

Variance and Covariance

The *variance* of $f(x)$ is

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

Covariance measures the extent to which two variables vary together.

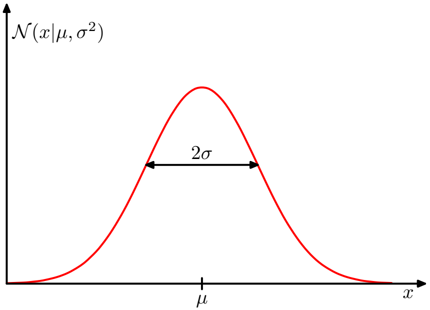
$$Cov(f(x), g(y)) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])(g(y) - \mathbb{E}[g(y)])]$$

The Gaussian Distribution

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

where

- μ is the mean, and
- σ is the standard deviation, where σ^2 is the variance.



Why the Gaussian is so widely used:

- Two easily interpretable parameters: mean and variance
- By Central Limit Theorem, sum of independent variables have $\sim$ Gaussian distribution
  - Makes a good choice for modeling noise
- Given a mean and variance, Gaussian makes least number of assumptions, i.e., has maximum entropy
- Simple mathematical form – easy to implement but usually highly effective

Closing Thoughts

- Probability is the mathematical foundation of machine learning.
- Loss functions will extend and build on this foundation, e.g. maximum likelihood estimation (MLE.)

KENNESAW STATE UNIVERSITY