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

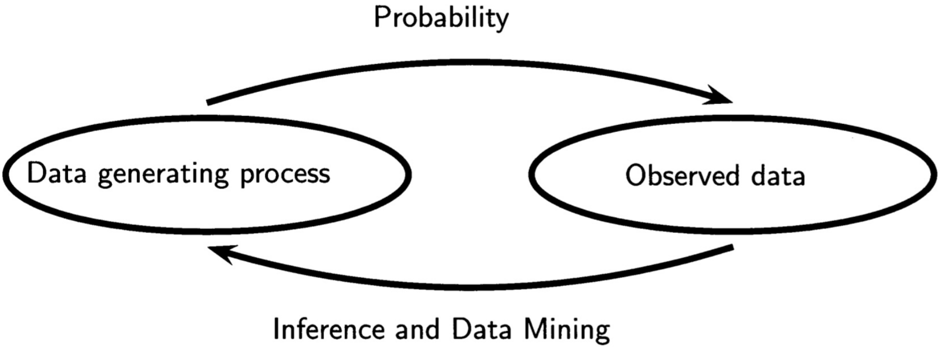
60%

40%

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

Probability in Machine Learning

We observe data generated by a random process.



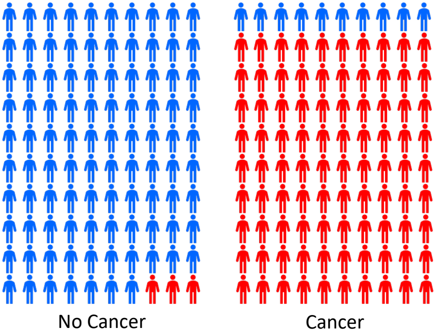
[2]

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

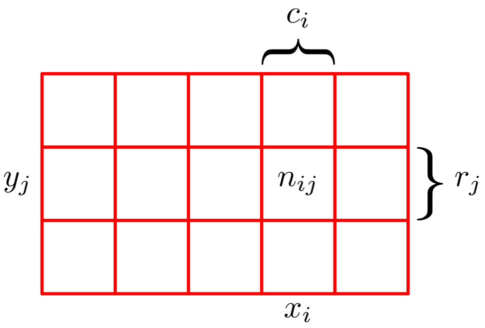
Let X and Y be *random* (a.k.a. *stochastic*) variables and

- $\{x_i\}_{i=1}^{L}$
- $\{y_j\}_{j=1}^{M}$
- 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

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

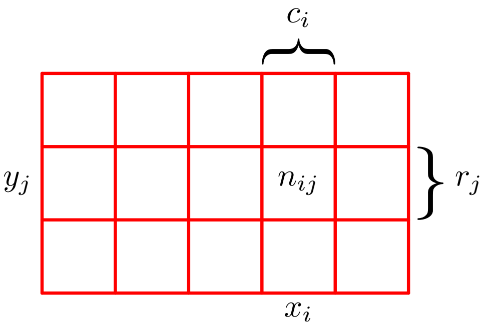
Notice that the number of instances in column i, c_i, is the sum of instances in each cell having n_{ij} instances, so $c_i = \sum_j n_{ij}$. Recalling that

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

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.



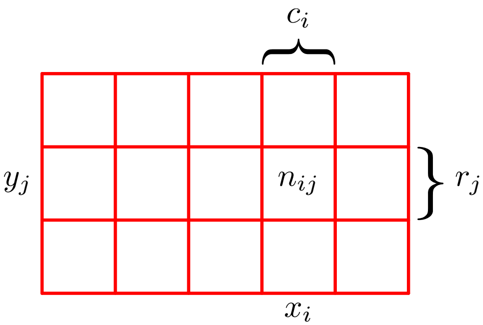
Conditional Probability

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:

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

This is the *product rule*.

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

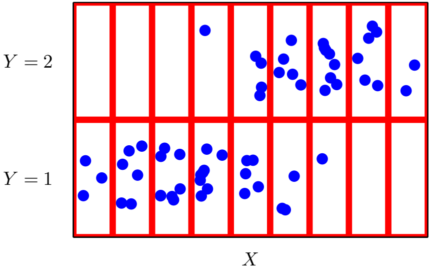
$$\text{Sum rule: } p(X) = \sum_Y p(X, Y)$$

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

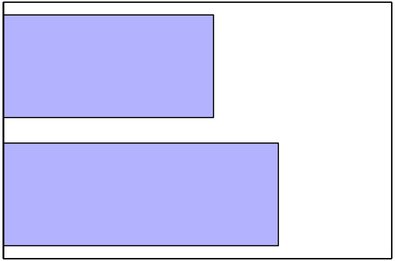
These two rules underlie all the probabilistic machinery we'll use in this course.

Visualizing Joint Distributions

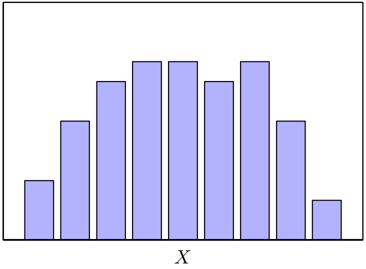
$p(X,Y)$



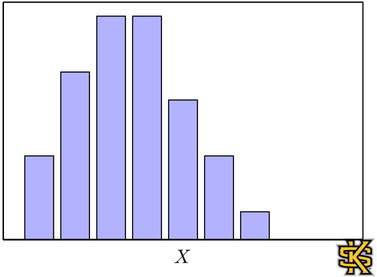
$p(Y)$



$p(X)$



$p(X|Y=1)$



Bayes' Theorem

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

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

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* (also called a posteriori probability) is our new belief after a Bayesian update.

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?

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?

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:

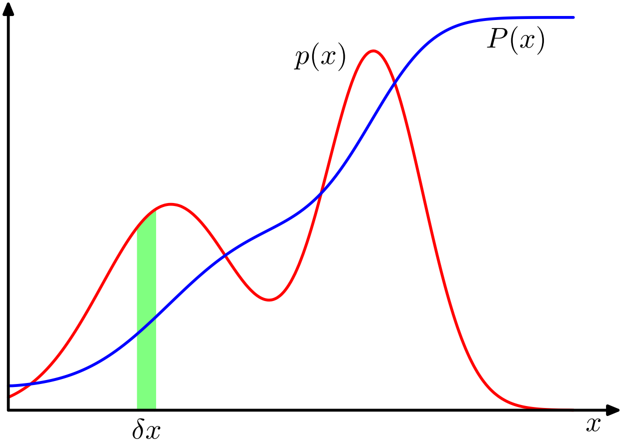
$$p(x \in (a, b)) = \int_a^b p(x)\, dx.$$

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

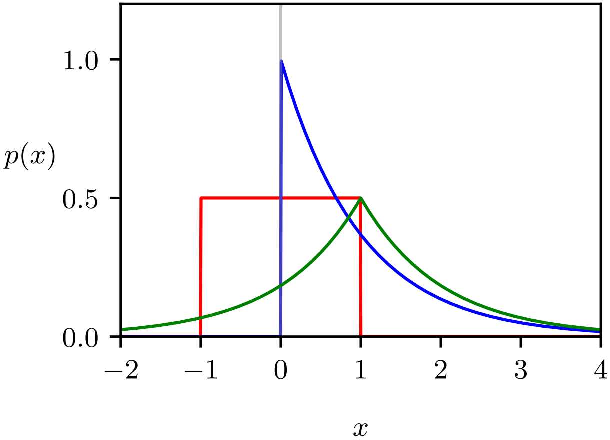
$$p(x) \geq 0$$

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

Probability Densities



Distributions



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

Uniform Distribution

2.2.1

Exponential Distribution

2.2.1

Lapace Distribution

2.2.1

Dirac Delta Function

2.2.1

Expectations

2.2.2

The *expected value* or *mean* or *forst moment* of a random variable X is the weighted average of the

Covariances

2.2.2

The Gaussian Distribution

2.3

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 – easily to implement but usually highly effective

The Gaussian Distribution

2.3

Mean and Variance

2.3.1

Likelihood Function

2.3.2

Maximum Likelihood

2.3.2

Why log:

- Log of a function monotonically increasing and concave – $\operatorname{argmax} ln(f) = \operatorname{argmax} f$
- Log easy to work with: $\ln(ab) = \ln(a) + \ln(b)$, $\ln(\frac{a}{b}) = \ln(a) - \ln(b)$
- Multiplying probabilities can underflow – summing logs avoids this problem

Bias of Maximum Likelihood

2.3.3

Linear Regression

2.3.4

Transformation of Densities

2.4

Maybe save for the Normalizing Flows lesson

Multivariate Distributions

2.4.1

Maybe save for the Normalizing Flows lesson

Information THeory

2.5

Entropy

2.5.1

Physics Perspective

2.5.2

Differential Entropy

2.5.3

Maximum Entropy

2.5.4

Kullback-Leibler Divergence

2.5.5

Conditional Entropy

2.5.6

Mutual Information

2.5.7

Bayesian Probabilities

2.6

Model Parameters

2.6.1

Regularization

2.6.2

Maximum Aposteriori (MAP) estimate

Bayesian Machine Learning

2.6.3

Closing Thoughts

Boom!