CSE 475: Statistical Methods in AI **Monsoon 2019**

SMAI-M-2019 4: Mathematical Foundations of ML - IV

Lecturer: C. V. Jawahar *Date: 8 Aug 2019*

4.20 Problem Space - IV

In the last lecture we looked at the learning problem and

- Understood it as an optimization problem of an appropriate loss/objective function.
- We also defined the data/examples $\mathcal{D}$ into two subsets $\mathcal{D}_{Tr}$ and $\mathcal{D}_{Te}$ as the subsets used for "Training" and "Testing"

Given this background, let us ask a critical question? What are we optimizing over?

- $\mathcal{D}_{Tr}$ or $\mathcal{D}_{Te}$ or $\mathcal{D}$?.

Since our objective is to define a computational procedure, we work only on $\mathcal{D}_{Tr}$.

This means our problem is something like:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{i \in \mathcal{D}_{Tr}} ||f(\mathbf{x}_i, \mathbf{w}) \leftrightarrow y_i||$$

Where $\leftrightarrow$ compares the predictions (i.e., $f(\mathbf{x}_i, \mathbf{w})$) with that of "truth" (i.e., y_i). It could be something as simple as difference.

The above formulation seems to be correct (and that is what we are going to use also!!). But there is a very important issue/puzzle.

If we use a simple Look Up Table (LUT), that stores $\mathbf{x}_i, y_i)$ and give you y_i for your ($\mathbf{x}_i$, does it minimize the above loss function? Yes; Indeed this gives zero error (perfect function) on the training set. However, a completely useless solution for the test set or for the purpose of learning. What went wrong in our problem formulation?

Really speaking, the problem we want to solve is slightly different, we would like to get a minimally complex function $f()$ that satisfy our data.

Occam's Razor: Prefer shorter hypothesis that fits the data.

- How do we define the "shortness"/complexity for the hypothesis/function $f()$.
- How do we find the shortest/shorter one?

The real problem of our interest is then

$$\min_{\mathbf{w}, f} \sum_{i \in \mathcal{D}_{Tr}} ||f(\mathbf{x}_i, \mathbf{w}) \leftrightarrow y_i|| + ||f||$$

where $||f||$ is a measure of complexity of $f()$. Without going to the details, we make certain notes here.

- The complexity of the function class $f()$ can be computed in different ways. (Say degree of a polynomial defines how complex is the polynomial. Or number of weights in a Neural network defines how complex is the neural network function.). In more formal machine learning literature, "VC" dimension is used as a complexity measure.
- Unfortunately, searching over a function class $f \in \mathcal{F}$ is not simple computationally. Searching/Trying out all neural networks and then picking the one that optimizes the above problem does not seem to be very attractive.
- In practice, we fix $f()$, and then look for $\mathbf{w}^*$. With experience you will pick a suitable function $f()$ and move forward to find the optimal $\mathbf{w}^*$. This also answers your worry about why are we trying out many solutions/hyperparameters.
- Also note that the optimization problem that we solve could be non-convex. This means that the chance of getting a good solution depends on how well we optimize and also what function class we pick.

4.20.1 Overfitting

We know that we split the data into training and testing and build the model based on the training data. There is a common serious pitfall, that is expected. The performance on training data could be very good. While performance on the test data could be very bad. This is popularly known as overfitting and should be avoided.

Overfitting Worry against overfitting is a serious concern among practitioners of machine learning. An over complex function class is more likely to overfit your training data (like LUT). This explains why do we need simple models that fit the data.

When performance of the algorithm is superior on training data and inferior on test data, we say that the algorithm is overfitting the training data.

Generalization: Though ML problems look to be very similar to the classical modelling/fitting problems with data, we always aim at doing well on the "unseen" data. Our performance is defined as the performance on the unseen data and not on the training data.

Generalization typically refers to a machine learning solution's ability to perform well on new or unseen samples rather than the training data or data that it has used/seen while trainig. It is also related to the concept of overfitting. If the model is overfitted, then it will not generalize well. We work hard to avoid overfitting.

4.21 Probabilistic View Point

Let us revisit our problem of classifying an email as spam or non-spam. It may be important to make a final classification as 0 or 1. However, in many situations what we would like to obtain is the probability of the email being spam or non-spam. This may be useful for situations like:

- a human to look closely and take the decision.
- our classification is any way under uncertainty. capture this uncertainty.
- results of this stage is used for many tasks in the subsequent stages.

Probabilistic view of the classification also allows us to incorporate the prior knowledge we have, along with the evidences we have to make optimal decisions. We will see some such examples later today.

Also a number of tools and techniques from statistics and probability theory helps us in formulating and interpretting the formulations solutions.

4.22 Terms and Definitions

We assume that a student of this course has gone through a basic course on probability theory. There are a number of terms you should recollect at this stage.

- Random Variables and Probability
- Probability Density Function
- Types of Probabilities
- Marginal Probability
- Conditional Probability
- Joint Probability
- Popular Distribution
- Normal Distribution
- Beta Distribution
- Popular Results
- Sum Rule of Probability
- Product Rule of Probability
- IID
- And many more

Do read a brief note on these associated concepts in the annexure at the end of this.

4.23 Bayes Theorem

4.23.1 Example

Let us start with an example of Bayes decision in discrete case.

You are captured by the *Sentinelese* tribe while on your excursion to the islands. You are brought to the chieftain for prosecution. You are blindfolded and the chief selects a fruit from a basket containing 85 green mangos, 5 yellow mangoes, 2 green pears and 8 yellow pears. If you guess the fruit correctly, you are set free. If not ..

- What is your guess?
- What is your chance of survival?

Simple Solution

$$P(\text{Mango}) = \frac{90}{100} = 0.9$$

$$P(\text{Pear}) = \frac{10}{100} = 0.1$$

So the safe bet is Mango. Isn't?

Evidence: Decisions are usualy not that simple. You will have more evidence to analyse the situation.

1. You get a glimpse through the blindfold and you see a slight yellow color in the chiefs hand.
2. Unfortunately you are colour-blind and you mistake green for yellow 20% of the time, but never yellow for green.

- What is your best guess?
- What will be your chance of survival now?

4.23.2 Bayes Theorem

Conditional probability : Conditional probability is the probability of observing an event, given the fact that a second event has occurred. Using formal notations, we write:
$P(fruit = mango/you\ saw\ yellow)$: read as $P(fruit = mango)$ given *you saw yellow*; or in short as: $P(mango/yellow)$.

Priori and posterior probabilities:

- Class prior probabilities $P(\omega_i)$
- In our example, this would be $P(mango)$ and $P(pear)$.
- The class-conditional probability density function $p(\mathbf{x}/\omega_i)$. The probability density function for x given the state of nature is ω_i
- In the example above the class conditional probabilities are $p(yellow/mango)$, $p(green/mango)$ etc.

Bayes Rule: Bayes rule states that the joint probability of $\mathbf{x}$ and ω_i, denoted as $p(\mathbf{x}, \omega_i)$ is given by:

$$p(\mathbf{x}, \omega_i) = p(\mathbf{x}/\omega_i).P(\omega_i) = P(\omega_i/\mathbf{x}).P(\mathbf{x})$$

We can rewrite the second equality as:

$$P(\omega_i/\mathbf{x}) = \frac{p(\mathbf{x}/\omega_i).P(\omega_i)}{P(\mathbf{x})}$$

Here the L.H.S is the posterior probability of class ω_i after observing $\mathbf{x}$. Bayes decision rule says to choose that ω_i which maximises the posterior probability. The above equation may also be written as:

$$P(\omega_i/\mathbf{x}) = \frac{p(\mathbf{x}/\omega_i).P(\omega_i)}{\Sigma_{j=1}^{c} p(\mathbf{x}/\omega_j).P(\omega_j)}$$

Bayes rules gives you a mathematical formula for combining the evidence (what you saw) with your prior knowledge (what you knew about number of fruits and their colours). The combined probability is usually called *posterior probability.*

Using Bayes rule we write:

$$P(mango/yellow) = \frac{p(yellow/mango).P(mango)}{P(yellow)}$$

Or

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

Bayes formula helps to convert the prior probability $P(\omega_i)$ to the *a posteriori* probability $P(\omega_i|x)$

Given the priors are equal, the category ω_j for which $p(x|\omega_j)$ is large is more *likely.*

If you are not colour blind:

$$P(\text{Mango/Yellow}) = \frac{\frac{5}{90} \cdot \frac{90}{100}}{\frac{5}{90} \cdot \frac{90}{100} + \frac{8}{10} \cdot \frac{10}{100}} = 0.385$$

$$P(\text{Pear/Yellow}) = \frac{\frac{8}{10} \cdot \frac{10}{100}}{\frac{5}{90} \cdot \frac{90}{100} + \frac{8}{10} \cdot \frac{10}{100}} = 0.615$$

Evidence can chnage your apriori decision!!

If you are colour blind:

$$P(\text{Mango/Yellow}) = \frac{\frac{5+0.2*85}{90} \cdot \frac{90}{100}}{\frac{5+0.2*85}{90} \cdot \frac{90}{100} + \frac{8+0.2*2}{10} \cdot \frac{10}{100}} = 0.724$$

$$P(\text{Pear/Yellow}) = \frac{\frac{8+0.2*2}{10} \cdot \frac{10}{100}}{\frac{5+0.2*85}{90} \cdot \frac{90}{100} + \frac{8+0.2*2}{10} \cdot \frac{10}{100}} = 0.276$$

4.24 Normal Distribution

We are familiar with the Normal/Gaussian distribution with mean μ and variance σ^2 from the school

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

In this case, x is the single variable. As we had seen in the problems of interest, our $\mathbf{x}$ is a vector consisting of $x_1, \ldots, x_d$. This naturally, demand the multivariate case as

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}[\mathbf{x}-\mu]^T\Sigma^{-1}[\mathbf{x}-\mu]\right)$$

Indeed when $d = 1$, both these equations become the same. Naturally, our mean will be a d dimensional vector. And the covariance Σ is a $d \times d$ matrix.

$$\mu = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$$

$$\Sigma = \frac{1}{N}[\mathbf{x}-\mu][\mathbf{x}-\mu]^T$$

- Q: What do the elements of Σ imply?
- Q: What are the properties of Σ?
- Q: How is this covariance matrix related to the correlation matrix ?
- Q: By looking at the covariance matrix, what all we can say?

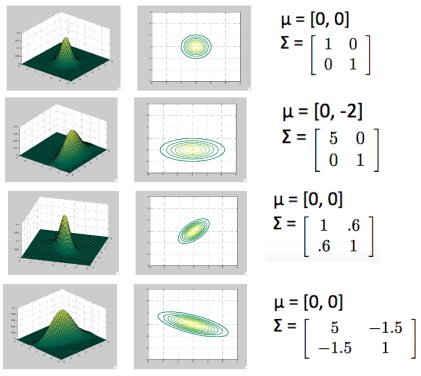


Figure 4.2: Appreciating the Covariance Matrix Structure

- Q: Why certain types of covariance matrices like $\Sigma = \sigma^2 I$ are of importance?
- We often model classes as multivariate Gaussians. Or we assume that there is an expected behaviour (measurement) for a class such as mean and there is a small deviation from the expected behavior that is modelled as Normal distribution.
- The quantity $[\mathbf{x}-\mu]^T\Sigma^{-1}[\mathbf{x}-\mu]$ is of special interest to us. This is called Mahalanobis distance.

CSC471: Statistical Methods in AI: Lecture Note 1
A Light Introduction to Random Variables and Density Functions

Anoop M. Namboodiri
IIIT, Hyderabad, INDIA. *anoop@iiit.ac.in*

1 Introduction

Random variables are the primary mechanisms by which one deals with variability, noise and uncertainty of real-world phenomena, their observations and inferences, in statistical pattern recognition (SPR).

Consider a variable x, which represents the height of a college student in India. Let us assume that the height of a students can be anywhere between $150\ cm$ and $190\ cm$. i.e, $x \in [150, 190]$. If we randomly select a student (draw a random sample), we will get a height between 150 and 190. Hence we call x, the height of a college student in India, a *random variable*, which assumes a specific value from its range, every time we draw a sample.

Let us assume that we conduct the following experiment: We randomly select a college student from India and measure his/her height. We can ask several questions regarding the outcome of our experiment.

- How likely are we to get a student of height, say 172?
- If we repeat the experiment 500 times, how many samples will have a height greater than 180?
- Are all the height measurements equally likely?
- If not, what is the most likely value for height?
- What is the expected value of height? Is it the same as above?
- What is the expected value of height, given that the gender of the sample is, say female?

By the end of this tutorial, you should be able to answer all the above questions (and those at the end) with clear reasoning. Specifically, the last question is most interesting from the point of view of pattern classification, which asks the inverse question, "what is the most likely gender of a sample, given that the height is 165?" Before we dive deeper into the details, we introduce a few terms that will be useful through the remainder of this tutorial.

The set of all possible samples in the problem is referred to as the *population*. This could be a finite set as in the case of 'all college students in India', or an infinite set, say all possible ways in which one can write the character *'a'*. Conversely, the unit that is selected from the population in each *trial* of the experiment is referred to as a *sample*. In our example, each college student is a sample. One might consider a different experiment where each trial involves randomly selecting a set of 10 students, and the random variable x is the number of different languages that they speak. Here, each sample would be 'a set of 10 students', and the 'set of all possible subsets of size 10' from the students forms the *sample space* of x.

The process of selecting a sample is called *sampling*. The sampling process can be repeated *with replacement* or *without replacement*, depending on whether a drawn sample is put back into the population before the next sample is drawn or not. In most cases, we make the following assumptions about the samples that are drawn from a sequence of trials:

1. *Independence:* The outcome of a particular trial (the sample that is drawn) has no bearing on the outcome of the following trials. i.e, the samples are independent of each other.
2. *Identical distribution:* The probability that any particular sample is drawn is the unchanged across the trials. In other words, the probability distribution is identical for all trials.

We put the two assumptions together and claim that the samples in an experiment are *independent and identically distributed* (*i.i.d.* or *iid* for short). The above assumptions are the primary reasons why we can make any inference about a population from a relatively small set of samples drawn from the population. We often assume that the method of sampling is *(simple) random sampling*, where every sample in the population has an

equal chance of being drawn in any trial. Note that for a finite population, applying random sampling with replacement will make the resulting samples, *iid*.

In the following sections, we deal with two different types of random variables: *discrete* and *continuous*. The distinction is based on the nature of values that a random variable can take. The tools required to deal with them might also be different, which will be discussed in detail.

2 Discrete Random Variables

In our initial example, if the heights of students are measured to the nearest centimeter, the set of values that x can take are $150, 151, \ldots, 190$. x will then be a discrete random variable (DRV).

A random variable x is called a *discrete random variable*, if the set of possible values that x can take is countable. The set of values are usually finite, although not necessarily so. A discrete random variable will always take one of the n values in its range, $\chi = \{v_1, v_2, \ldots, v_n\}$ (for now, we assume n to be finite).

2.1 Probability *Mass* Function (PMF)

Now, let us consider one of our initial questions: In our random draw experiment, 'How likely are we to get a student of height 172? If our sampling is random, then the probability of drawing a sample of height 172 depends only on the number of students having height 172 in the population. Let the number of students having a height h is n_h out of the total population of N students. The probability that a randomly selected student has height h is n_h/N, as every sample has an equal probability of getting selected.

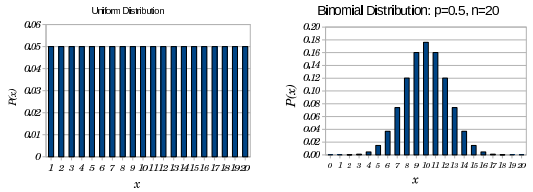
In general, the probability that a discrete random variable, x takes a value v_i (i.e., $Pr[x = v_i]$) is denoted as p_i. We denote the function that maps each value $v_i \in \chi$ to its occurrence probability, p_i, as $P(v_i)$. As we saw, the probability p_i can be computed as the fraction of the population with a value of $x = v_i$. The function $P(.)$ can be thought of as representing the distribution of the population over the values in χ, and hence is called the *Probability Distribution Function* or the *Probability Mass Function (PMF)*. To avoid confusion with similar terms, we will call $P(.)$ as the probability mass function or *PMF* in the case of discrete random variables.

As noted above, each value of p_i is a probability and the PMF should satisfy the following conditions:

$$\forall_i, P(v_i) \geq 0, and \tag{1}$$

$$\sum_{i=1}^{n} P(v_i) = \sum_{i=1}^{n} p_i = 1. \tag{2}$$

Certain parametric forms of the probability mass function are popular in practice, as they model the process of generation of the samples. Figure 1 shows two popular PMF forms, uniform and binomial.



(a) (b)

Figure 1. (a) uniform and (b) binomial distributions for PMF.

A discrete random variable is completely characterized by its PMF. Any other property of the random variable can be derived completely and precisely from its PMF. In other words, we can find the answers to all the questions posed in the introduction, if we can compute the PMF of the random variable, 'height of a college student in India'!!, well, almost all. We will now look into two of the important properties of a random variable, its expectation and variance.

2.2 Expectation and Variance: μ & σ^2

The expected value of a discrete radom variable, x is defined as:

$$\mathcal{E}[x] \equiv \mu = \sum_{x \in \chi} xP(x) = \sum_{i=1}^{n} v_i P(v_i). \tag{3}$$

What does this expectation tell us? The expected value is a *weighted average* of all possible values of x, weighted by their probabilities. In other words, μ is just the mean value of x over the entire population. Here are a couple of other ways in which we can think about μ.

- Assume that each sample in the population has unit mass, and is placed in space according to the value of x, v. The expected value, μ, will give you the centre of mass of the whole population.
- If we are asked to guess the outcome of the experiment over a large number of trials, and if we guess μ every time, we will make the least error, overall, in the MSE sense. That is why we call it the expected value.

However, if we were to guess the most likely height among all students, we will be better off guessing the mode of the distribution and not its mean . . . ofcourse!!

Now we know the best guess of the outcome of our experiment. However, can we say anything about the amount of error we will make? This is precisely what the variance tells us.

The *variance*, σ^2 of a random variable is defined as:

$$\text{Var}(x) \equiv \sigma^2 = \mathcal{E}[(x-\mu)^2] = \sum_{i=1}^{n}(v_i-\mu)^2 P(v_i). \quad (4)$$

As you can see, the variance is the mean squared error (MSE) if you guess the mean. If the *mean* tells you about the centre of mass of the population, the *variance* tells you how spread out the population is from the mean. Note that variance only gives you a measure of spread of the data and not the exact way in which it is spread. For that you need the complete PMF itself.

One can also represent the variance as:

$$\begin{aligned}
\sigma^2 &= \mathcal{E}[(x-\mu)^2] = \int_{-\infty}^{\infty}(x-\mu)^2 p(x)dx \\
&= \int (x^2 - 2x\mu + \mu^2)p(x)dx \\
&= \int x^2 p(x)dx - 2\mu\int xp(x)dx + \mu^2\int p(x)dx \\
&= \mathcal{E}[x^2] - 2\mu\mathcal{E}[x] + \mu^2.1
\end{aligned}$$

which simplifies to:

$$\sigma^2 = \mathcal{E}[x^2] - (\mathcal{E}[x])^2. \quad (5)$$

Note: If we compute the square root of the variance, i.e, RMSE w.r.t μ, we get the *standard deviation*, σ.

3 Continuous Random Variables

In the previous section, we assumed that the height measurement of a student is an integer value, making x a discrete random variable. If we assume that the height can be measured precisely to any real number between 150 and 190, the number of values that x can take will become uncountable. Such random variables, which usually take any value within a continuous range are referred to as *continuous random variables* (CRV). Note that the number of real numbers in a range are uncountable. In each trial, the random variable, x, can take any of the infinite number of values within its range, χ. The range could also be infinite, i.e. $(-\infty, \infty)$.

In our example, even though the range is finite ($[150, 190]$), the number of possible values that x can take are uncountably infinite. This makes the definition of probabilities, tricky.

3.1 Probability *Density* Function (PDF)

Consider the continuous domain equivalent of the first question that we asked: 'If we randomly select a student, how likely are we to get a specific height, say 172.3413587391, precise up to the picometer or more?' Intuitively, we can say that it is extremely unlikely, well almost impossible, that we will chance upon a student with that exact height. i.e, $Pr[x = 172.3413587391...] = 0$. Then what about exactly $172.00...$? or any other specific real number between 150 and 190? We have to say they also have the same plight. To generalize, the probability that a continuous random variable takes any specific value in its range is 0. Does that mean no event can ever occur?!!

To get around this predicament, we reframe the question a bit as follows: 'How likely are we to select a student of height within the range $[172-\delta, 172+\delta]$? Now there is a non-zero probability that we might get a number within that range. Based on this, we define the distribution of samples in the range as follows:

For every continuous random variable, x, there exists a *probability density function*, $p(x)$, such that:

$$\forall_x, p(x) \geq 0, \quad and \quad (6)$$

$$Pr[x \in (a,b)] = \int_a^b p(x)dx. \quad (7)$$

From the second condition, we can also infer that:

$$\int_{-\infty}^{\infty} p(x)dx = 1 \quad (8)$$

$p(x_t)$ gives the limiting value for density of probability in a small window around the point x_t. Note that the value of $p(x_t)$ is not a probability. We always use lower case p for densities, and upper case P for functions that give probabilities. The probability density function (PDF), plays the same role for CRVs as PMF for DRVs. Note that in theory the PDF need not be a parametric function, although in practice it always is.

3.2 Expectation and Variance: μ & σ^2

We can extend the definitions of expected value and variance of a RV from the discrete to continuous domain as the following integrals:

$$\begin{aligned}
\mathcal{E}[x] &\equiv \mu &&= \int_{-\infty}^{\infty} xp(x)dx \\
\text{Var}(x) &\equiv \sigma^2 &&= \int_{-\infty}^{\infty}(x-\mu)^2 p(x)dx.
\end{aligned}$$

These measures also have similar meanings or interpretations as we found for the discrete RVs. To make the ideas clear, we will consider two examples of PDFs:

3.2.1 Uniform Density

The uniform density function is characterized by the range within which it is defined as is given by:

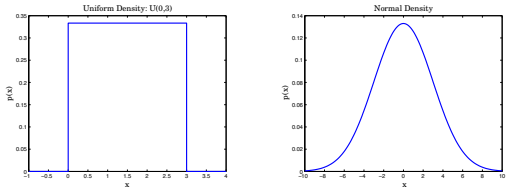
$$U(a,b) = \begin{cases} \frac{1}{(b-a)} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

$$\begin{aligned} \mu &= \int_a^b x \left(\frac{1}{b-a}\right) dx = \frac{1}{b-a}\int_a^b x dx \\ &= \frac{1}{b-a}\left[x^2/2\right]_a^b = (b+a)/2, \end{aligned}$$

which is what we expect of the mean of a uniform distribution between a and b. Similarly, the variance can shown to be:

$$\sigma^2 = \frac{(b-a)^2}{12} \quad (10)$$

Figure 2(a) shows the plot of a uniform density function in the range $[0, 3]$.



(a) (b)

Figure 2. (a) uniform and (b) norma densities for PDF.

3.2.2 Normal Density

The Normal or Gaussian density is one of the most popular density functions in practice, as it is a good approximation of many real world random processes. The normal density function, $N()$ has two parameters, μ, and σ, and is given by:

$$N(\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (11)$$

Figure 2(b) shows a normal density plot: $N(0, 3)$, within the range $[-10, 10]$. Note that the support of a normal density if infinite. The expectation and variance of the normal density function are in fact, μ and σ^2 themselves.

In addition to what we discussed, there are a large number of probability distributions for both discrete and continuous RVs that are used in specific scenarios [1].

4 CDF: Cumulative Distribution Function

The cumulative distribution function or CDF is derived from the PDF by the integral of the density up to a point. It is defined as:

$$C(t) = \int_{-\infty}^{t} p(x)dx. \quad (12)$$

Not that the CDF gives the total probability that a continuous random variable takes a value less than a specific value, t. The CDF is can be expressed in a parametric form in certain cases, such as the uniform density:

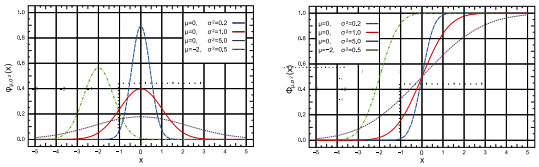
$$C(t) = \begin{cases} 0 & \text{if } t < a \\ \frac{(t-a)}{(b-a)} & \text{if } a \leq t \leq b \\ 1 & \text{if } t > b \end{cases} \quad (13)$$

Note that a PDF of a RV completely specifies its CDF and vice-versa. However, it is possible that one of them has a compact parametric representation, while the other does not. For example, the CDF of the normal distribution (equation 11) is given by:

$$cdf(x) = \frac{1}{2}\left(1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right), \quad (14)$$

where $erf()$ is the error function defined by:

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_{-\infty}^{x} e^{-t^2} dt. \quad (15)$$



(a) (b)

Figure 3. (a) normal densities with different parameters and (b) the their CDFs [1].

There is no closed form representation to the erf, and it is often approximated by its Taylor series expansion. Figure 3 shows the normal density function with four different parameters, and the corresponding cumulative distribution functions.

4.1 Generating Random Numbers

One of the very useful applications of CDFs is that one can generate random numbers that follow any given distribution, provided we can compute/estimate the CDF of the distribution.

Consider a random variable, x that is distributed according to a PDF, $p(x)$. Also consider another random variable, $y = C(x)$, where $C(x)$ is the CDF corresponding to $p(x)$.

Now, consider a small window of x around the point t, $[t - \delta t, t + \delta t]$ (see Figure 4). The value of y corresponding to t will be $r = C(t)$. Moreover, the value of y corresponding to $x = t+\delta t$ will be $r+\delta t.p(t)$, assuming that δt is small, giving $p(t + \delta t) \approx p(t)$. Similarly, $C(t - \delta t) = r - \delta t.p(t)$.

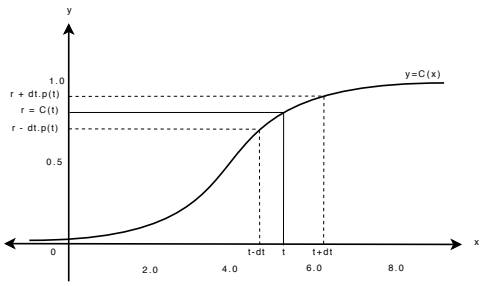


Figure 4. Mapping of a random variable using the CDF.

In other words, all samples of x within a window of size $2\delta t$ around t will map to a window of size $2\delta t.p(t)$ around $C(t)$ for y. The resulting density of y will be hence $1/p(t)$ times the density of x, which is unity. i.e, y is of uniform density in the range $[0, 1]$.

We just argued that given a random variable x of any density, the corresponding random variable, $y = C(t)$ will be $U[0, 1]$. We can invert this statement and say that given a random variable y that follows the pdf $U[0, 1]$, the random variable $x = C^{-1}(y)$ will follow a PDF with corresponding CDF as $C()$. In other words given a set of random numbers y_i with uniform density $U(0, 1)$, we can map it to a set of random variables x_i with any desired PDF using the inverse CDF function !!!

5 Problems

1. Give an example each of probability mass functions with finite and infinite ranges. Show that the conditions on PMF are satisfied by your example.
2. Show with complete steps that the variance of uniform density is given by equation 10. (*Hint: use the expression for variance in equation 5.*)
3. Show examples of two density functions (draw the function plots) that have the same mean and variance, but clearly different distributions. Plot both functions in the same graph with different colours.
4. Show that the alternate expression for variance given in equation 5 holds for discrete random variables as well.
5. Prove that the mean and variance of a normal density, $N(\mu, \sigma^2)$ are indeed its parameters, μ and σ^2.
6. Using the inverse of CDFs, map a set of $10,000$ random numbers from $U[0, 1]$ to follow the following pdfs:
   (a) Normal density with $\mu = 0$, $\sigma = 3.0$.
   (b) Rayleigh density with $\sigma = 1.0$.
   (c) Exponential density with $\lambda = 1.5$.

   Once the numbers are generated, plot the normalized histograms (the values in the bins should add up to 1) of the new random numbers with appropriate bin sizes in each case; along with their pdfs. What do you infer from the plots? *Note: see rand() function in C for $U[0, INT\_MAX]$.*
7. Write a function to generate a random number as follows: Every time the function is called, it generates 500 new random numbers from $U[0, 1]$ and outputs their sum.

   Generate $50,000$ random numbers by repeatedly calling the above function, and plot their normalized histogram (with bin-size = 1). What do you find about the shape of the resulting histogram?

References

[1] *Probability distribution*, Wikipedia, 2008, http://en.wikipedia.org/wiki/Probability_distributions.

[2] R. Duda, P. Hart, and D. Stork, *Pattern classifcation and scene analysis*, 2^{nd} ed., John Wiley and Sons, New York, 2001.

[3] John A. Rice, *Mathematical statistics and data analysis*, 2nd ed., Duxbury Press, 1995.