A Refresher on Probabilities and K-means Clustering

Morteza Haghir Chehreghani

March 18, 2011

Overview

A Refresher on Probabilities

K-means Clustering

Refresher

Pen and Paper

Random Variables

- A random variable is a "probabilistic" outcome of an experiment, such as a coin flip or the height of a person chosen from a population.

- Notation:

  X Random variable
  $\approx$ a device from which we draw a value.

  x If x is not capital, it denotes a value taken by the RV X.
  $Pr\{X = x\}$ denotes the probability for this to occur.

  $\mathcal{X}$ Sample space or domain of X.
  The set of all values a draw from X may result in.

Random Variables

RVs take on values in a sample space.

Types of sample spaces:

1. Discrete sets:
    - Finite: for a coin flip $\mathcal{X} = \{H, T\}$
    - Infinite: $\mathcal{X} = \mathbb{N}, \mathbb{Z}$ etc.
2. Continuous sets: e.g. $\mathcal{X} = \mathbb{R}, \mathbb{R}_+, \mathbb{R}^d, [0,1], [a,b]$

Probability distribution function describes how probabilities are distributed over the values of the random variable:

- $p(\mathrm{x}) =$ the probability that X takes the value x.

Probability of Random Variables

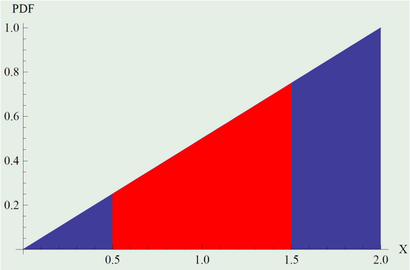
- A discrete distribution assigns a probability to every atom in the sample space of a random variable.
- For example, if X is an (unfair) coin, then the sample space consists of the atomic events $X = H$ and $X = T$, and the discrete distribution might look like:
  $P(X = H) = 0.7$
  $P(X = T) = 0.3$
- For any valid discrete distribution, the probabilities over the atomic events must fulfill:
  1. Non-negativity: $P(x) \geq 0$
  2. Normalization: $\sum_{x \in \mathcal{X}} P(x) = 1$

Continuous Random Variables

- A continuous random variable can assume any value in an interval or in a collection of intervals.

$$P(a \leq X \leq b) = \int_a^b p(x)dx$$

Example: Find the probability that $0.5 \leq X \leq 1.5$



Continuous Random Variables

- For continuous probability distributions, we require:
  1. Non-negativity: $p(x) \geq 0$
  2. Normalization: $\int_{\mathcal{X}} p(x) dx = 1$
- **Notation:** We deal with three types of symbols:

$\Pr\{...\}$	Probability of an event (inside the curly brackets), such as $\Pr\{X = x\}$.
P(x)	Probability mass function.
p(x)	Probability density function.

- Density functions are only applicable in the case of continuous sample spaces.

Joint Probabilities

Typically, one considers collections of RVs.
For example, the flipping of 4 coins involves 4 RVs, 1 for each coin.

Joint probability:	The probability for precisely the values x, y to occur together.
Definition:	$P(x,y) := \Pr\{X = x, Y = y\}$

The joint distribution for a flip of each of 4 coins assigns a probability to every outcome in the space of all possible outcomes of the 4 flips.

If all coins are fair:

$$
\begin{aligned}
P(HHHH) &= 0.0625 \\
P(HHHT) &= 0.0625 \\
P(HHTH) &= 0.0625 \\
&\ldots
\end{aligned}
$$

Conditional Probability

A conditional distribution is the distribution of some random variable given some evidence, such as the value of another random variable.

- **Def.**: $P(X = x|Y = y)$ is the probability that $X = x$ when $Y = y$.

Conditional probability can be defined in terms of the joint and single probability distributions:

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

(which holds when $P(Y) > 0$)

The Chain Rule

The definition of conditional probability leads to the chain rule, which lets us define the joint distribution of two (or more) random variables as a product of conditionals:

The Chain Rule:

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

- The chain rule can be used to derive the $P(X, Y)$ when it is not known.
- The chain rule can be extended to any set of n variables.

Marginalization

- Given a collection of random variables, we are often interested in only a subset of them. For example, we might want to compute $P(X)$ from a joint distribution $P(X, Y, Z)$.

Def.

Marginal probability:	The probability for x to occur, regardless of y.
Discrete case:	$P(x) := \sum_{y \in \mathcal{Y}} P(x, y)$
Continues case:	$p(x) := \int_{\mathcal{Y}} p(x, y) dy$

Marginalization

This property actually derives from the chain rule:

$$
\begin{aligned}
\sum_{y \in \mathcal{Y}} P(x,y) &= \textstyle\sum_{y \in \mathcal{Y}} P(x)P(y|x) && \text{by the chain rule} \\
&= P(x) \textstyle\sum_{y \in \mathcal{Y}} P(y|x) && P(x) \text{ doesn't depend on y} \\
&= P(x) && \textstyle\sum_{y \in \mathcal{Y}} P(y|x) = 1
\end{aligned}
$$

Bayes Rule

By the chain rule:

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

This is equivalently expressed as Bayes rule:

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

Independence

- Random variables are independent if knowing about X tells us nothing about Y. That is,

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

.

- This means that their joint distribution factorizes:

$$P(X, Y) = P(X)P(Y)$$

- This factorization is possible because of the chain rule:

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

I.i.d

- I.i.d. $=$ independently, identically distributed
- RVs $X_1, ..., X_n$ are i.i.d. iff
  1. They are (pairwise) statistically independent.
  2. All drawn according to the same distribution.

- Note: If $X_1, ..., X_n$ are i.i.d., then

$$\begin{aligned} p(x_1, ..., x_n) &= p_{X_1}(x_1)...p_{X_n}(x_n) \\ &= \prod_{i=1}^{n} p(x_i) \end{aligned}$$

Expectation

- Definition:
$$\mu_x := \mathrm{E}[X] := \int_{\mathcal{X}} x p(x) dx$$
The integral is called the first moment of p.
- Note: Expected value $\neq$ Most likely value.
- For a function f:
$$\mathrm{E}[f(X)] := \int_{\mathcal{X}} f(x) p(x) dx$$

Variance

- Definition:

$$\sigma_X^2 := \mathrm{Var}[X] := \int_{\mathcal{X}} (x - \mu_X)^2 p(x) dx$$

$\rightarrow$ second centralized moment of p.

- Always: $\mathrm{Var}[X] \geq 0$
- Definition: The square root $\sigma_X = \sqrt{\mathrm{Var}[X]}$ is called the standard deviation of X.

Statistics

- Expectation and variance map distribution functions (densities or mass functions) to real values. They are examples of functionals of distribution functions.
- Note: A functional is a mapping which takes a function as its argument.
- Definition: We call a functional of a distribution function a statistic of the distribution.

Multiple Dimensions

- A vector of random variables

$$\mathbf{X} = (X_1, ..., X_n)^\top$$

  A draw $\mathbf{x} = (x_1 \dots x_n)^\top$ from $\mathbf{X}$ defines a point in n-dimensional space.
- It is treated just like a list of $1D$ RV's.
- The vector components are not necessarily i.i.d
- We can add RV's to produce a new RV

$$Y := c_1 X_1 + c_2 X_2$$

Multidimensional Moment Statistics

- Expectation: Vector of components expectation

$$\mathrm{E}[\mathbf{X}] := (\mathrm{E}[X_1], ..., \mathrm{E}[X_n])^\top$$

- Variance: Generalized to covariance:

$$\begin{aligned} Cov[X, Y] &:= \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y)(x - \mu_X)(y - \mu_Y) dx dy \\ &= \mathrm{E}_{X,Y}[(x - \mu_X)(y - \mu_Y)] \end{aligned}$$

- If X,Y are independant, then $Cov[x, y] = 0$
- Proportional behavior:

$$\begin{aligned} Cov[X, Y] > 0 &\Leftrightarrow X, Y \text{increase together} \\ Cov[X, Y] < 0 &\Leftrightarrow X, Y \text{are anti-proportional} \end{aligned}$$

Covariance Matrix

- For RVs $X_1, ..., X_n$ we use a covariance matrix Σ to describe their mutual covariances:

$$\Sigma_{i,j} := Cov[X_i, X_j] \qquad i, j = 1, ..n$$

Properties:

1. Diagonal entries are RVs variances

$$\Sigma_{i,j} := Cov[X_i, X_i] = Var[X_i]$$

2. Σ is symmetric

$$\Sigma_{i,j} = Cov[X_i, X_j] = Cov[X_j, X_i] = \Sigma_{j,i}$$

3. Σ is positive semi-definite

Brain Teaser

Question: Assume you have observed 2D data $\mathbf{X} \in \mathbb{R}^{2 \times N}$ (observations as columns). The first row of $\mathbf{X}$ corresponds to the first dimension x_1, the second row corresponds to x_2.

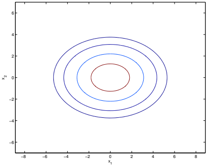
x_1	1.5	4.3	...	0.2
x_2	2.7	-2.1	...	6.0

For each of the 3 covariance matrices $\mathbf{C_X}$, choose the iso-line plot (A-E) corresponding to the covariance matrix.

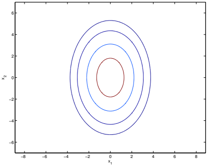
Brain Teaser

1. $\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ 2. $\begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$ 3. $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

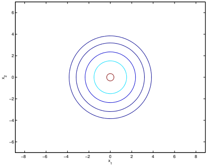
A



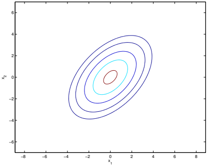
B



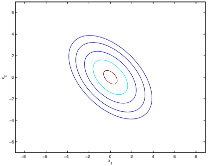
C



D

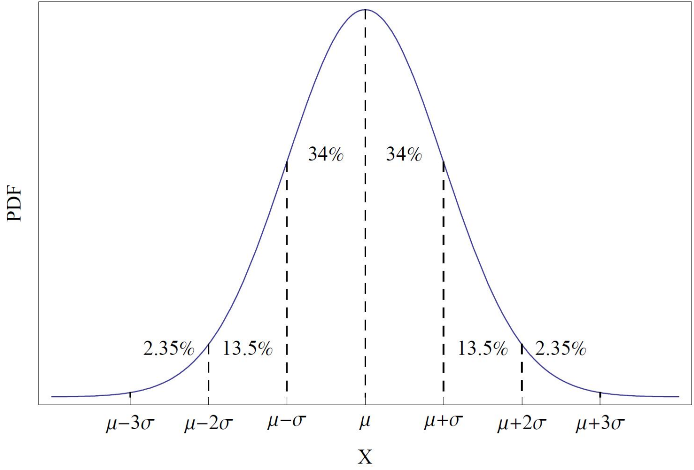


E



Gaussian Distribution (1D)

- Sample space $\mathcal{X} = \mathbb{R}$
- Definition: $p(x|\mu, \sigma) := \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$



Gaussian Distribution (nD)

- Sample space $\mathcal{X} = \mathbb{R}^n, \mathbf{x} = (x_1, .., x_n)^\top$
- Definition:
  $p(\mathbf{x}|\mu, \Sigma) := \frac{1}{(\sqrt{2\pi})^n |\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\mathbf{x} - \mu)^\top \Sigma^{-1} (\mathbf{x} - \mu))$

  where Σ is the covariance matrix and $|\Sigma|$ is its determinant

Data vs. Distribution

- Important: Be careful to distinguish between distributions (smooth functions in most examples) and data (point clouds).
- Machine learning:
  - Data = input
  - Distribution = model or assumption
- ML methods usually make some general assumptions about distribution, then try to obtain ("infer") the specifics from the data.

Example 1) Modeling step: Assume a Gaussian as model.
2) Inference step: Estimate Gaussian parameters (μ and σ) from data.

Empirical distribution

- We try to regard data sample (imagine some point cloud) as a distribution.
- Problem: We only know wether or not a point is there, not how probable that is.
- Simple solution: Assign same probability to each point.

  Def. Let $S = \{x_1, .., x_n\}$ be a sample of the data, we call

  $$P(x) := \frac{1}{n} \cdot \#\{y \in S | y = x\}$$

  the empirical distribution defined by the data.

The Clustering Problem

- Consider N data points in a D-dimensional space. Each data vector is denoted by $\mathbf{x}_n$, $n = 1, \ldots, N$.
- Our goal is to partition the data set into K clusters.
- In other words, find vectors $\mathbf{u}_1, \ldots, \mathbf{u}_K$ that represent the centroid of each cluster.
- A data point $\mathbf{x}_n$ belongs to cluster k if the Euclidean distance between $\mathbf{x}_n$ and $\mathbf{u}_k$ is smaller than the distance to any other centroid.

K-means Cost Function

Objective

Minimize the following cost function

$$J(\mathbf{U}, \mathbf{Z}) = \|\mathbf{X} - \mathbf{U}\mathbf{Z}\|_2^2 = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{k,n} \|\boldsymbol{x}_n - \boldsymbol{u}_k\|_2^2.$$

Here, $\mathbf{X} = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_N] \in \mathbb{R}^{D \times N}$, $\mathbf{U} = [\mathbf{u}_1 \ \cdots \ \mathbf{u}_K] \in \mathbb{R}^{D \times K}$. We call the $\mathbf{u}_k$ the centroids. And $\mathbf{z}_n$ the assignments of data points to clusters.

Constraints on $\mathbf{Z}$: Hard assignments

We consider the constraint $\mathbf{Z} \in \{0, 1\}^{K \times N}$ with $\sum_k z_{k,n} = 1 \ \forall n$, i.e., one element per column set to 1.

K-means Algorithm

1. Initiate with a random choice of $\mathbf{u}_1^{(0)}, \ldots, \mathbf{u}_K^{(0)}$ (or let $\mathbf{u}_1^{(0)}, \ldots, \mathbf{u}_K^{(0)}$ equal data points from the set), set $t = 1$.
2. **Cluster assignment.** Solve $\forall n$:

$$k^*(\mathbf{x}_n) = \underset{k}{\operatorname{argmin}} \left\{ \|\mathbf{x}_n - \mathbf{u}_1^{(t)}\|_2^2, \ldots, \|\mathbf{x}_n - \mathbf{u}_k^{(t)}\|_2^2, \ldots, \|\mathbf{x}_n - \mathbf{u}_K^{(t)}\|_2^2 \right\}.$$

   Then, $z_{k^*(\mathbf{x}_n),n}^{(t)} = 1$ and $z_{j,n}^{(t)} = 0 \ \forall j \neq k$, $j = 1, \ldots, K$.
3. **Centroid update.** The centroids are given by:

$$\mathbf{u}_k^{(t)} = \frac{\sum_{n=1}^{N} z_{k,n}^{(t)} \mathbf{x}_n}{\sum_{i=1}^{N} z_{k,n}^{(t)}} \ \forall k, \ k = 1, \ldots, K$$

4. Increment t. Repeat step 2 until $\|\mathbf{u}_k^{(t)} - \mathbf{u}_k^{(t-1)}\|_2^2 < \epsilon \ \forall k$ $(0 < \epsilon \ll 1)$ or until $t = t_{\text{finish}}$.

K-means Exercise

Question 1: Show that the K-means algorithm always converges.

Hint: Show that both steps only decrease the objective, unless the algorithm converged.

K-means Exercise

Question 2: Formally show that the K-means Algorithm can be recast as a Matrix Factorization problem.

Again, the cost function

$$J(\mathbf{U}, \mathbf{Z}) = \|\mathbf{X} - \mathbf{U}\mathbf{Z}\|_2^2 = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{k,n} \|\boldsymbol{x}_n - \boldsymbol{u}_k\|_2^2,$$

under the constraint $\mathbf{Z} \in \{0,1\}^{K \times N}$ with $\sum_k z_{k,n} = 1$.

- Show that at **Step 2**, for a given $\mathbf{u}$, the K-means algorithm solves:

$$\min_{\mathbf{Z}} \sum_{n=1}^{N} \sum_{k=1}^{K} \|\mathbf{x}_n - z_{k,n}\mathbf{u}_k\|_2^2$$

- Show that at **Step 3**, for a given $\mathbf{Z}$, the K-means algorithm solves:

$$\min_{\mathbf{u}} \sum_{n=1}^{N} \sum_{k=1}^{K} \|\mathbf{x}_n - z_{k,n}\mathbf{u}_k\|_2^2$$