Gaussian mixture models unit:

- Everything you need to know about probability (**today**)
- Lab activity: From histograms to Gaussians (next week Wednesday)
- Mixture models (next week Friday)
- Expectation-maximization (week-after Monday)

Today:

- Overview of unit
- Definition of discrete probability
- Discrete random variables
- Continuous random variables and distributions
- The Gaussian distribution

Given (scalar) observations $\boldsymbol{x}$ generated by a process suspected of being comprised of $K$ subprocesses, each with a Gaussian distribution, train a model to predict the probability of observation value $x$, using the following model family:

$$p(x) \text{ or } p(x, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{i=0}^{K-1} \pi_i \mathcal{N}(x \mid \mu_i, \sigma_i)$$

where $\pi_i$ is the probability of an observation having come from subprocess $i$ and $\mu_i$ and $\sigma_i$ are the mean and standard deviation, respectively, of subprocess $i$, and $\mathcal{N}$ is the probability density function for the Gaussian distribution,

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

That is, find $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, and $\boldsymbol{\sigma}$ to maximize the likelihood of the training data under this model.

Probability is a way to model the potential outcomes of an experiment.

- Pulling a card from a deck... which card?
- Flipping a coin... heads or tails?
- Rolling a die (or two dice)... how many dots facing up?
- Dropping a Tbsp of cookie dough on a baking sheet... how many chocolate chips?
- Examining an iris... what length petals?

The set of basic outcomes in the experiment is the *sample space*.

- Each of 52 cards
- $\{H, T\}$
- {⚀, ⚁, ⚂, ⚃, ⚄, ⚅}
- $\mathbb{W}$
- $\mathbb{R}$

An *event* is a set of basic outcomes from the sample space.

- $\{\heartsuit J, \diamondsuit J, \clubsuit J, \spadesuit J\}$
- { ⚃⚅, ⚅⚃, ⚄⚄ }
- $\{0, 1, 2, 3\}$
- $[3, 4)$

Let $\Omega$ be a sample space and $\mathcal{F} = \mathscr{P}(\Omega)$ be an event space; A *probability function* $P : \mathcal{F} \to [0,1]$ fulfills the axioms of probability:

1. For all $A \in \mathcal{F}$, $P(A) \geq 0$.
2. $P(\Omega) = 1$
3. For disjoint sets $A, B \in \mathcal{F}$, $P(A \cup B) = P(A) + P(B)$.

Consider the events

- $P(\{\heartsuit J, \heartsuit Q, \heartsuit K, \diamondsuit J, \ldots \clubsuit K\}) = \frac{12}{52} \approx .23$
- $A$, the card is red. $P(A) = .5$
- $B$, the card is a diamond. $P(B) = .25$
- $C$, the card is a 4. $P(C) = \frac{4}{52} \approx .077$
- $D$, the card is $\diamondsuit 4$. $P(D) = \frac{1}{52} \approx .019$

A real random variable provides us with a numerical value that is dependent on the outcome of an experiment. It is a convenient way to express the elements of $\Omega$ as numbers rather than abstract elements of sets. Throughout this book, we will only consider **real** random variables or **multivariate real** random variables, that is to say, random variables with values in $\mathbb{R}$ or $\mathbb{R}^d$ for $d \geq 2$.

**Definition 2.4.1.** A real random variable $X$ is a function $X : \Omega \to \mathbb{R}$ such that for all $B \in \mathscr{P}(\mathbb{R}), \{\omega \in \Omega \mid X(\omega) \in B\} \in \mathcal{F}$.

The above definition naturally extends to $X : \Omega \to \mathbb{R}^d$ for all $d \geq 2$.

Han Veiga and Ged, pg 17

In machine learning, we often avoid explicitly referring to the probability space, but instead refer to probabilities on quantities of interest, which we denote by $\mathcal{T}$. In this book we refer to $\mathcal{T}$ as the *target space*.

We introduce a function $X : \Omega \to \mathcal{T}$ that takes an outcome and returns a particular quantity of interest $x$ as a value in $\mathcal{T}$. This association/mapping from $\Omega$ to $\mathcal{T}$ is called a *random variable*

The name "random variable" is a great source of misunderstanding as it is neither random nor is it a variable. It is a function.

Deisenroth et al, *Mathematics for Machine Learning*, pg 155

*Remark.* The target space, that is, the [codomain] $\mathcal{T}$ of the random variable $X$, us used to indicate the kind of probability space, i.e., a $\mathcal{T}$ random variable. When $\mathcal{T}$ is finite or countably infinite, this is called a discrete random variable. For continuous random variables, we consider only $\mathcal{T} = \mathbb{R}$ or $\mathcal{T} = \mathbb{R}^D$.

ibid, pg 157

A *random variable* is a function from an event space $\Omega$ to $\mathbb{R}$. A *discrete random variable* is a random variable whose range is a countable subset of $\mathbb{R}$.

The *probability mass function* $p_X$ of a discrete random variable $X$ is

$$p_X(x) = P(X = x) = P(\{\omega \in \Omega \mid X(\omega) = x\})$$

The *probability density function* $f_X$ of a continuous random variable $X$ is the function such that

$$P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt$$

or

$$P(a \leq X \leq b) = \int_{a}^{b} f_X(x)dx$$

Suppose we observe 10 whole-shell peanuts with the following number of seeds in each:

$$2, 2, 1, 2, 3, 2, 2, 3, 3, 2$$

Let $X$ be the random variable standing for the number of seeds in a peanut. The range of $X$ is $\{1, 2, 3\}$.

| Number of seeds | Probability |
|---|---|
| 1 | $\frac{1}{10}$ |
| 2 | $\frac{3}{5}$ |
| 3 | $\frac{3}{10}$ |

The *expected value* of a discrete random variable $X$ with probability density function $p_x$ is

$$\mathbb{E}[X] = \sum_x x \ p_x(x)$$

The expected value of a continuous random variable $X$ with probability mass function $p_x$ is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \ p_X(x) \ dx$$

The *expected value* of a continuous random variable $X$ with probability mass function $p_x$ is

$$\mu = \mathbb{E}[X] = \int_{-\infty}^{\infty} x \; p_X(x) \; dx$$

The *variance* of a random variable $X$ is the average distance from average squared.

$$\sigma^2 = Var(X) = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

The *standard deviation*, $\sigma$, is the square root of the variance.

A random variable $X$ has a *Gaussian distribution* if it has a probability density function in the form of

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

**Coming up:**

**Due Wed, Feb 19:**
*Do logistic regression programming assignment*

**Due Thurs, Feb 20:**
*Read from Chapter 2 in the textbook, about probability*
*(See Canvas for details)*
*Also, there will be a quiz (not ready yet)*

**Due Fri, Feb 21:**
*Submit "Dataset" checkpoint for term project*