10-707: Advanced Deep Learning **Spring 2020**

Lecture 9: February 17

Lecturer: Andrej Risteski *Scribes: Cinnie Hsiung, Euxhen Hasanaj, Weyxin Ly, Xinhe Zhang*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications, if as reader, you find an issue you are encouraged to clarify it on Piazza. They may be distributed outside this class only with the permission of the Instructor.*

9.1 Review of Lecture 8

9.1.1 Unsupervised Learning

Learning from data **without** labels.

What we can hope to do:

1. Structure Learning: Fit a parametrized structure to data to reveal something meaningful about data.
2. Distribution Learning: Learn a distribution close to data generating distribution.
3. Representation/Feature Leaning Learn a distribution that implicitly reveals an "embedding"/"representation" of data for downstream tasks.

9.1.2 Structure Learning

Methods of structure learning: Principal Component Analysis (PCA) and Clustering.

9.1.3 Principal Component Analysis (PCA)

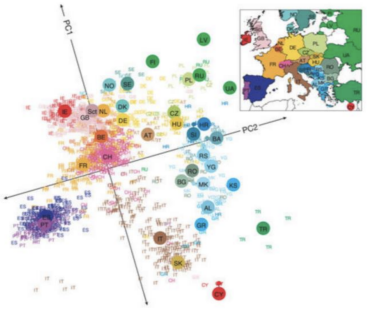


Figure 9.1: Example: Recover map of Europe from genome data with PCA.[NJB+08]

The Classics: PCA find a k-dimensional (linear) subspace explaining most of the variance in the data.

$$\max_{v_1,v_2,\ldots,v_k} \frac{1}{m} \sum_{\textbf{samples } x_i} (\text{length of } x_i \text{ on span}(v_1, v_2, \ldots, v_k))^2$$

$$= \max_{\text{orthonormal}\{v_1,v_2,\ldots,v_k\}} \frac{1}{m} \sum_{\text{samples } x_i} \sum_{j=1}^{k} \langle x_i, v_j \rangle^2 \leftarrow \text{ Variance: } \mathbb{E}\left[\Sigma_j \langle v_j, x \rangle^2\right]$$

9.1.4 K-means Clustering

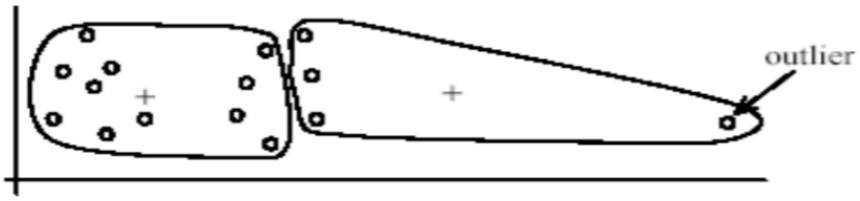
If the distance metric is the **Euclidean distance**, and the measure of cohesion is the **average distance from the centroid**: we get the *k-means objective*.

$$argmin_{\{r_{nk},\mu_k\}} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|X_n - \mu_k\|^2$$

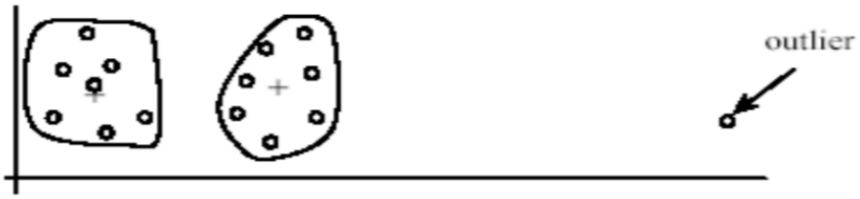
Where r_{nk} measures whether the point n is in cluster k, and μ_k is the centroid of k-th clustre.

Weaknesses of K-means clustering

Very sensitive to outliers (mean is very easily influenced by extreme values)

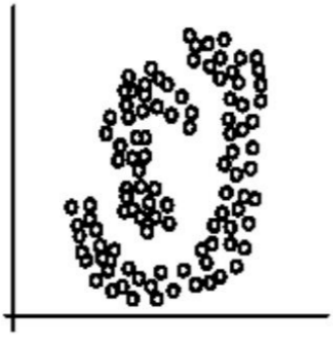


(A): Undesirable clusters

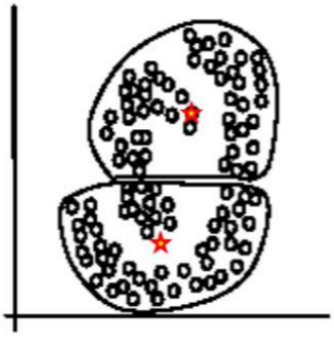


(B): Ideal clusters

Not suitable for non-spherical clusters



(A): Two natural clusters



(B): k-means clusters

9.2 Gaussian Distributions

9.2.1 Gaussian Univariate Distribution

The pdf of the **Gaussian distribution** is given by

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}, \tag{9.1}$$

where μ, σ^2 are the mean and variance respectively.

Theorem 9.1 *The mean $\mathbb{E}[X]$ of a random variable $X \sim \mathcal{N}(x, \mid \mu, \sigma^2)$ is given by μ.*

Proof: Without loss of generality (by change of variables), we can assume $\mu = 0$ and $\sigma = 1$. Then

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x \cdot \left(\frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \right) dx \\
&= \frac{1}{2} \int_{-\infty}^{\infty} (x + (-x)) \cdot \left(\frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \right) dx \\
&= 0 = \mu
\end{aligned}$$

where in the second equality we used the fact that the distribution is symmetric around 0, hence, the probability distributions for x and $-x$ are the same. ■

Theorem 9.2 *The variance $Var[X]$ of a random variable $X \sim \mathcal{N}(x, \mid \mu, \sigma^2)$ is given by σ^2.*

Proof: Without loss of generality (by change of variables), we can assume $\mu = 0$ and $\sigma = 1$. Then

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x^2 \cdot \left(\frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \right) dx \\
&= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} x \left(x \exp(-x^2/2) \right) dx \\
&= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} x dv \\
&= \frac{2}{\sqrt{2\pi}} \left(xv \Big|_0^{\infty} - \int_0^{\infty} v dx \right) \\
&= \frac{2}{\sqrt{2\pi}} \left(\int_0^{\infty} e^{-x^2/2} dx \right) \\
&= \frac{1}{\sqrt{2\pi}} \left(\int_{-\infty}^{\infty} e^{-x^2/2} dx \right) = 1
\end{aligned}$$

where we used the change of variables $v = -e^{-x^2/2}$ and hence $dv = x \exp(-x^2/2)$. ■

9.2.2 Multivariate Gaussian Distribution

For a d-dimensional vector x, the Gaussian distribution takes the form

$$\mathcal{N}(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right\},$$

where $\mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d}$ are the mean and covariance matrix respectively. Note also that $|\Sigma|$ denotes the determinant of Σ.

Theorem 9.3 *The mean $\mathbb{E}[X]$ of a random variable $X \in \mathbb{R}^d$, $X \sim \mathcal{N}(x, \mid \mu, \Sigma)$ is given by μ.*

Proof: Without loss of generality (by change of variables), we can assume $\mu = 0$. Then

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x \cdot \left(\frac{1}{(2\pi)^{d/2}} \frac{1}{|\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} x^T \Sigma^{-1} x \right\} \right) dx \\
&= \frac{1}{2} \int_{-\infty}^{\infty} (x + (-x)) \cdot \left(\frac{1}{(2\pi)^{d/2}} \frac{1}{|\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} x^T \Sigma^{-1} x \right\} \right) dx \\
&= 0 = \mu,
\end{aligned}$$

by the fact that the Gaussian is an even function. ■

Theorem 9.4 *The variance $Var[x]$ of a random variable $x \in \mathbb{R}^d$, $x \sim \mathcal{N}(x, \mid \mu, \Sigma)$ is given by Σ.*

Proof: Let $x' \sim \mathcal{N}(\mu, I)$ where I is the identify matrix. Then $x = \Sigma^{1/2} x'$. Then covariance of x is then given by

$$\begin{aligned}
x &= \mathbb{E}[xx^T] \\
&= \Sigma^{1/2} \mathbb{E}[x'(x')^T] \Sigma^{1/2},
\end{aligned}$$

it suffices to show that the covariance of x' is I.

We know that a Gaussian with $\Sigma = I$ is just a product distribution of standard Gaussians. Hence, the covariance matrix has diagonals 1 and off diagonals 0. ■

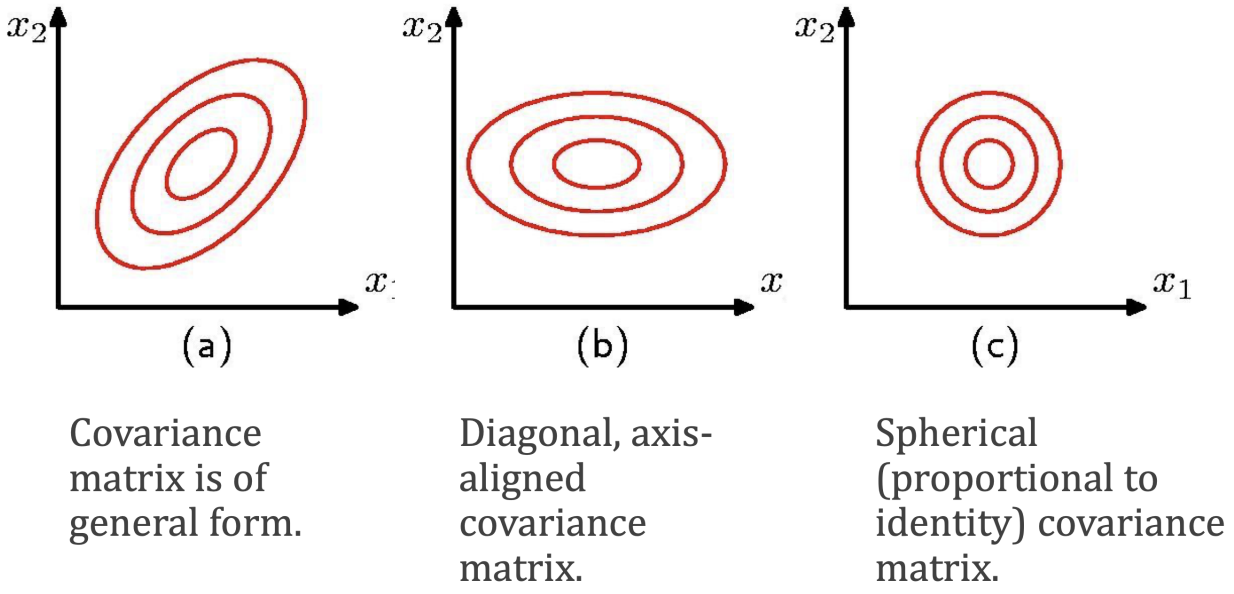


Figure 9.2: Example:Level sets of the probability densities for Gaussian RVs of different covariances.

9.2.3 Central Limit Theorem

One of the reasons why the Gaussian distribution is so popular is the Central Limit Theorem (CLT). Roughly, the CLT says that given N independent and identically distributed random variables, their sum becomes increasingly Gaussian as N grows. For example, assume we are given N uniformly distributed random variables in $[0, 1]$. Their mean

$$\frac{x_1 + \ldots x_N}{N}$$

tends towards a Gaussian distribution as N increases (can be made quantitative). See Figure 9.3. Perhaps more surprisingly, this holds for any distribution.

9.2.4 Marginals and Conditionals of Gaussians

Consider a d-dimensional Gaussian distribution $p(x) = \mathcal{N}(x \mid \mu, \Sigma)$. Consider the partitioning x into x_A and x_B where

$$x = \begin{bmatrix} x_A \\ x_B \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB}. \end{bmatrix}$$

We can define also the **precision matrix** $\Lambda \equiv \Sigma^{-1}$.

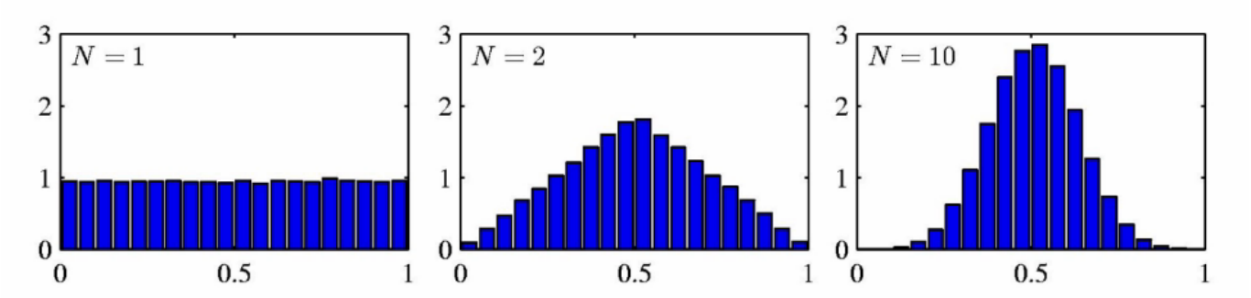


Figure 9.3: Plots of the distribution of the mean of N uniformly distributed random variables as N increases.

Lemma 9.5 *The **block inversion lemma** states that*

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix} \tag{9.2}$$

Theorem 9.6 *The conditional distribution of a Gaussian distribution is also a Gaussian distribution. That is,*

$$p(x_B \mid x_A) = \mathcal{N}(x_B \mid \mu_{B|A}, \Sigma_{B|A})$$

Proof: Let us start by writing down the conditional explicitly using Bayes rule

$$\begin{aligned} p(x_B|x_A) &= \frac{p(x_A, x_B; \mu, \Sigma)}{\int_{x_A \in \mathbb{R}^m} p(x_A, x_B; \mu, \Sigma) dx_A} \\ &= \frac{1}{Z'} \exp\left(-\frac{1}{2} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix}^T \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}^{-1} \begin{bmatrix} x_A - \mu_A \\ x_B - \mu_B \end{bmatrix} \right) \end{aligned}$$

where Z' is a normalizing constant. If we look at the term inside the exponential, this is going to be a quadratic in x_A so it looks like a Gaussian. Let's make this precise by massaging the expression using the matrix analogue of "completing the square". Using the same notation as before for the precision matrix $\Lambda = \Sigma^{-1}$, and expanding the block form, we get

$$\begin{aligned} p(x_B|x_A) = \frac{1}{Z'} \exp\Big(-\Big[&\frac{1}{2}(x_A - \mu_A)^T \Lambda_{AA}(x_A - \mu_A) + \frac{1}{2}(x_A - \mu_A)^T \Lambda_{AB}(x_B - \mu_B) \\ &+ \frac{1}{2}(x_B - \mu_B)^T \Lambda_{BA}(x_A - \mu_A) + \frac{1}{2}(x_B - \mu_B)^T \Lambda_{BB}(x_B - \mu_B) \Big] \Big) \\ = \frac{1}{Z'} \exp\Bigg(-\Bigg[&\frac{1}{2}\Big(x_B - \mu_B + \Lambda_{BB}^{-1}\Lambda_{BA}(x_A - \mu_A)\Big)^T \Lambda_{BB} \Big(x_B - \mu_B + \Lambda_{BB}^{-1}\Lambda_{BA}(x_A - \mu_A)\Big) \\ &+ \frac{1}{2}(x_A - \mu_A)^T \Lambda_{AA}(x_A - \mu_A) - \frac{1}{2}(x_A - \mu_A)^T \Lambda_{AB}\Lambda_{BB}^{-1}\Lambda_{BA}(x_A - \mu_A) \Bigg] \Bigg). \end{aligned}$$

Now we shall use the following "completing the square" trick

$$\frac{1}{2} z^T A z + b^T z + c = \frac{1}{2}(z + A^{-1}b)^T A (z + A^{-1}b) + c - \frac{1}{2} b^T A^{-1} b.$$

For simplicity, let's denote $z = x_B - \mu_B, A = \Lambda_{BB}, b = \Lambda_{BA}(x_A - \mu_A), c = 1/2(x_A - \mu_A)^T \Lambda_{AA}(x_A - \mu_A)$. With this new notation, and using the fact that $\Lambda_{BA}^T = \Lambda_{AB}$, we have

$$p(x_B|x_A) = \frac{1}{Z'} \exp\left(-\left[\frac{1}{2}(z + A^{-1}b)^T A (z + A^{-1}b) + c - \frac{1}{2} b^T A^{-1} b \right] \right).$$

The last terms can be grouped into the normalizing constant and it is clear in this form that the conditional is also Gaussian, with mean μ_B and variance Σ_{BB} (i.e., we just drop the corresponding rows/columns).

■

Theorem 9.7 *The marginal distribution of a Gaussian distribution is also a Gaussian distribution. That is*

$$p(x_A) = \int p(x_A, x_B) dx_B = \mathcal{N}(x_A \mid \mu_A, \Sigma_{AA}).$$

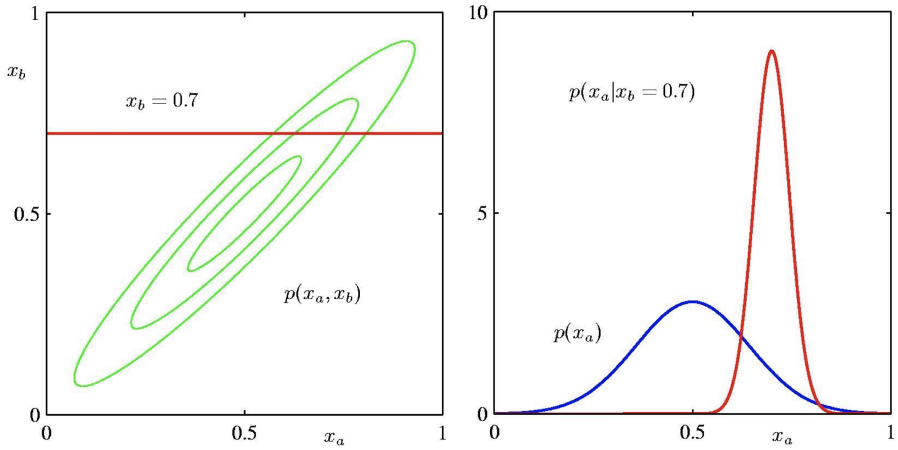


Figure 9.4: Example: Graphical depiction of the conditional and marginal distribution of a Gaussian.

9.2.5 Mixture of Gaussians

Naive Gaussian distributions may not be sufficient to model all real world data. Instead, we can combine simple models into a complex model by adding K Gaussian densities in the form

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k),$$

where π_k is the mixing coefficient and the Gaussians (which each have their own mean and variance) are the mixture components.

Note that more generally, mixture models can comprise of linear combinations of other distributions.

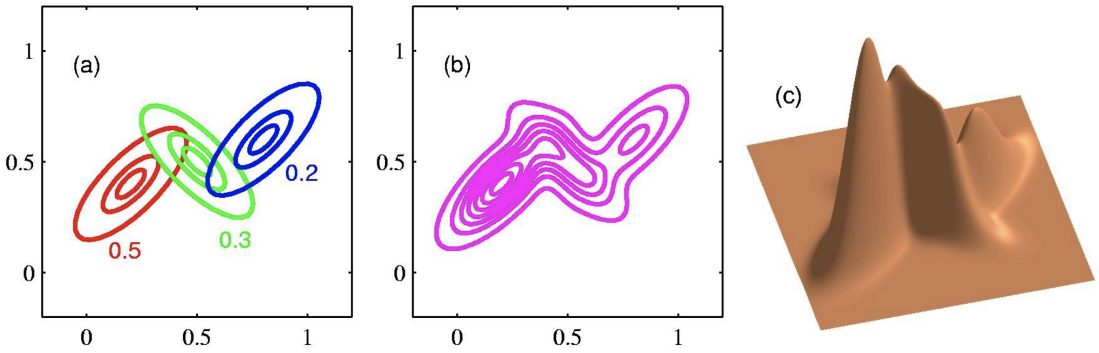


Figure 9.5: Example: Illustration of a mixture of 3 Gaussians in a 2D. The subfigures represent (a) contours of constant density of each of the mixture components along with the mixing coefficients, (b) contours of probability density $p(x)$, and (c) a surface plot of the distribution $p(x)$.

9.3 Bernoulli Distribution

Bernoulli Distribution: Discrete probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability $1 - p$

Consider a single binary random variable $x \in \{0, 1\}$, such as the outcome of flipping a coin. Let us denote the value 1 as getting a heads and the value 0 as getting a tails.

The probability of $x = 1$ can be denoted by the parameter μ so that:
$p(x = 1|\mu) = \mu \qquad 0 \leq \mu \leq 1$

As a result, the Bernoulli distribution can be written as:
$\text{Bern}(x|\mu) = \mu^x(1 - \mu)^{1-x}$

The **expected value** of a Bernoulli random variable is:
$E[x] = 1 * P(x = 1) + 0 * P(x = 0) = \mu$

The **variance** of a Bernoulli random variable is:
$Var[x] = E[x^2] - E[x]^2 = \mu - \mu^2 = \mu(1 - \mu)$

9.4 Binomial Distribution

The binomial distribution can be thought of as the distribution which tracks the outcome of multiple coin tosses. In this sense it can be seen as the probability distribution of a sum of Bernoulli random variables. More concretely, assume $X \sim \text{Bernoulli}(\mu)$, then the probability of observing m heads given N independent coin flips is given by

$$\mathbb{P}(m \text{ heads}|N, \mu) = \text{Bin}(m|N, \mu) = \binom{N}{m}\mu^m(1 - \mu)^{N-m}.$$

This can be derived by noticing that in N coin tosses, we want m of them to be heads, and the rest to be tails which contributes the $\mu^m(1-\mu)^{N-m}$ part, however, we should take into account the number of all such possible combinations which is precisely $\binom{N}{m}$ and that gives the above formula.

The mean and the variance for the binomial can be easily derived by using the fact that it is just a sum of N Bernoulli's each with mean μ and variance $\mu(1-\mu)$. Therefore,

$$\mathbb{E}(m) = N\mu$$
$$\text{Var}(m) = N\mu(1-\mu).$$

For large N (a rule of thumb is $N > 30$ and p not close to 0 or 1), the binomial distribution can be approximated by a normal distribution $\mathcal{N}(N\mu, N\mu(1-\mu))$. This follows by the law of large numbers.

9.5 Categorical Distribution

Categorical Distribution: Discrete probability distribution that describes the possible results of a random variable that can take on one of K possible categories, with the probability of each category separately specified and the probabilities of each possible outcome must be in the range from 0 to 1 and all must sum to 1.

Consider rolling a dice that can take on $K = 6$ states. If we observe a roll that corresponds to state 3, then $x_3 = 1$. Then, we can represent x using a 1-of-K encoding scheme (a vector with one element containing 1 and all other elements containing a 0) as: $x = (0, 0, 1, 0, 0, 0)^T$. In this form, the categorical distribution is equivalent to a multinomial distribution for a single observation.

Let us denote the probability of observing state k as $x_k = 1$.
Then, the categorical distribution over x can be defined as:
$P(x|\mu) = \prod_{k=1}^{K} \mu_k^{x_k} \quad \forall \mu_k \geq 0 \text{ and } \sum_{k=1}^{K} \mu_k = 1$

The categorical distribution is a generalization of the Bernoulli distribution for more than 2 outcomes.
$P(x|\mu) = \prod_{k=1}^{K} \mu_k^{x_k}$

Additionally, it is easy to see that the distribution is normalized.
$\sum_x p(x|\mu) = \sum_{k=1}^{K} \mu_k = 1$
$E[x|\mu] = \sum_x P(x|\mu)x = (\mu_1, ..., \mu_k)^T = \mu$

9.6 Multinomial Distribution

Multinomial Distribution: A generalization of the binomial distribution

For example, consider modeling the probability of counts for each side of a k sided dice rolled n times. Each of the n independent trials results in a success in exactly one of the k categories, with the probabilities of each possible category being in the range from 0 to 1 and all must sum to 1. The multinomial distribution gives the probability of any particular combination of numbers of successes for the k various categories.

We can construct the joint distribution of the quantities $\{m_i\}$ given the parameters $\{\mu_i\}$ and the total number of observations:
$\text{Mult}(m_1, m_2, ..., m_k|\mu, N) = \binom{N}{m_1 m_2 ... m_k} \prod_{k=1}^{K} \mu_k^{m_k}$

The expected number of times the outcome k was observed over N trials is:
$E[m_k] = N\mu_k$

Each diagonal entry of the covariance matrix is the variance of a binomially distributed random variable.

$Var[m_k] = N\mu_k(1-\mu_k)$

The off diagonal entries of the covariance matrix are: $Cov[m_j m_k] = -N\mu_j\mu_k$
All covariances are negative because for fixed n, an increase in one component of a multinomial vector requires a decrease in another component.

The normalization coefficient is the number of ways of partitioning N objects into K groups of size $m_1, m_2, \ldots, m_K$.

9.7 Graphical Model

Graph contains a set of nodes connected by edges.

In a **probabilistic graphical model**, each node represents a random variable, links represent "probabilistic dependencies" between random variables.

Graph specifies how joint distribution over all random variables **decomposes** into a **product** of factors, each factor depending on a subset of the variables.

There are two types of graphical models:

1. **Bayesian networks**, also known as **Directed Graphical Models**, where the links have a particular directionality indicated by the arrows.

2. **Markov Random Fields**, also known as **Undirected Graphical Models**, where the links do not carry arrows and have no directional significance.

9.7.1 Bayesian Networks

Directed Graphs are useful for expressing causal relationships between random variables.

Example: Given your symptoms are: fever + red spots, what's the probability that you have measles?

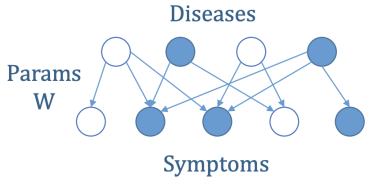


Figure 9.6: Example: Disease and symptom bayesian network.

The Bayesian network succinctly describes $Pr[symptom|diseases]$.

9.7.1.1 Noisy-OR Networks

Given $d_i, s_i \in \{0, 1\}$ and $W_{ij} \geq 0$:

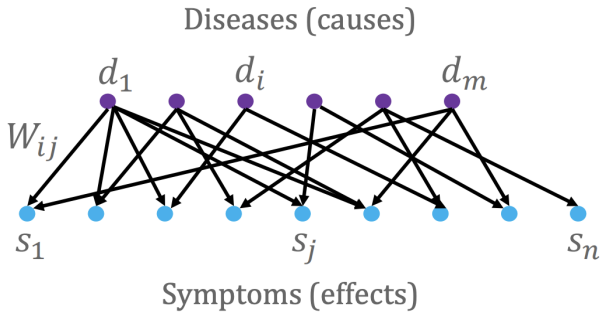


Figure 9.7: Example: Disease and symptom noisy-or network

This network is noisy because each d_i is on **independently** with probability p. When d_i is on, it **activates** s_j with probability $1 - exp(-W_{ij})$. And s_j is **on** if one of d_j's **activates** s_j.

9.7.1.2 "Deriving" Bayesian Networks as restrictions of arbitrary distributions

An **arbitrary** joint distributions $p(a, b, c)$ over three random variables a,b, and c can be written as

$$p(a, b, c) = p(c|a, b)p(a, b) = p(c|a, b)p(b|a)p(a)$$

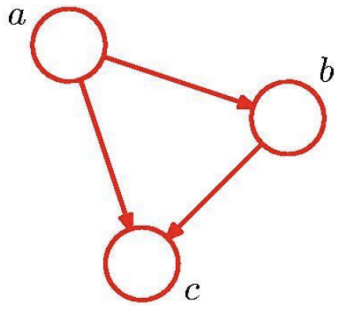


Figure 9.8: Example: Associate graph with distributions.

To associate a graph with the decomposition:

1. Node for each of the random variables.
2. Add directed links to the graph according to the distribution.

As a result, different ordering can result to different graphical representation.

Joint distribution over K variables factorizes: $p(x_1, \ldots, x_K = p(x_K|x_1, \ldots, x_{K-1} \ldots p(x_2|x_1)p(x_1)$. And the corresponding undirected graph is fully connected.

A graph that is **not** fully connected conveys information about the conditional **independence** structure of the distribution it encodes.

Example:

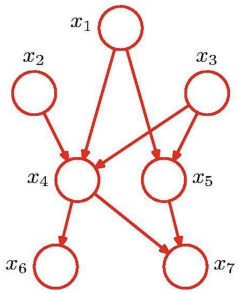


Figure 9.9: Example: Not fully conected graph conveys independence information.

The graph above encode distributions over $x_1, \ldots, x_7$ that can be written as:

$$p(x_1, \ldots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$

The graph is not fully connected at x_5 since it is only conditional on x_1 and x_3.

The joint distribution defined by the graph is given by the product of a conditional distribution for each node **conditioned on its parents**: $p(x) = \Pi_{k=1}^{K} p(x_k|pa_k)$.

Restriction: There must be **no directed cycles**. (i.e., the graph is a Directed Acyclic Graph (DAG))

9.7.1.3 Crucial Property: Easy Sampling

One advantage of bayesion network is easy sampling.

Consider the graph in the example above, suppose each of the conditional distributions are easy to sample from. Then how do we sample from the joint?

We can start at the top and sample in order.

$\hat{x_1} \sim p(x_1)$
$\hat{x_2} \sim p(x_2)$
$\hat{x_3} \sim p(x_3)$
$\hat{x_4} \sim p(x_4|\hat{x_1}, \hat{x_2}, \hat{x_3}) \leftarrow$ The parent variables are set to their sampled values
$\hat{x_5} \sim p(x_5|\hat{x_1}, \hat{x_3})$

To obtain a sample from the marginal distribution, e.g. $p(x_2, x_5)$, sample from the full joint distribution, retain $\hat{x_2}$, $\hat{x_5}$, and discard the remaining values.

9.7.1.4 Typical Deep Learning Application

Higher-up nodes will typically represent **latent** (hidden) random variables.

Latent variables model distributions over observed variables constructed from simpler conditional distributions.

Example: Latent-variable model of image.

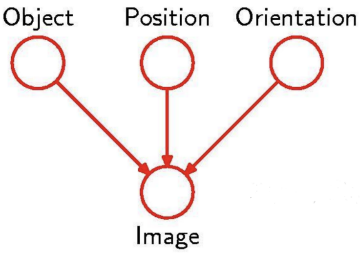


Figure 9.10: Example: Latent-variable model of image

In this example, object identify, position and orientation have independent *prior probabilities.*

Image has probability distribution that depends on these three properties.

Therefore,

$$P(Im, Ob, Po, Or) = P(Im|Ob, Po, Or)P(Ob)P(Po)P(Or)$$

Where $P(Im|Ob, Po, Or)$ is the *likelihood* and $P(Ob)P(Po)P(Or)$ is the *prior.*

Likelihood and prior are modeled by parametric distribution whose parameters are fitted throughout training.

9.7.2 Why restrict connectivity

1. Fully connected graphs restrict the richness of the class.
2. For a discrete joint distribution over **n** variables, where each variable takes one of **k** values. To fully specify it in general, we need to specify k^n values.
3. But to specify the conditionals $p(x_1|x_2, x_3, \ldots, x_d)$ we need to specify $k^d \ll k^n$ values.

References

[NJB+08] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al. Genes mirror geography within europe. *Nature*, 456(7218):98–101, 2008.