CISC 271 Class 36

The Curse of Dimensionality

Text Correspondence: Zaki *et al.* [19], pp. 163–182

Main Concepts:

- *Some problems grow exponentially with the number of dimensions*
- *Labeling problems can grow super-exponentially*
- *Uniformly distributed data are mainly at hypercube "edges"*
- *Normally distributed data are mainly in the "tails"*

Sample Problem, Data Analytics: How do problems "grow" with the number of dimensions?

In the preface (p. ix) of Richard Bellman's 1956 technical report [1], the founder of dynamic programming wrote that a solution is

> *". . . quite definitely not routine when the number of variables is large"*

A more famous sentence and phrase of Bellman's, on the same page, is:

> *"All this may be subsumed under the heading 'the curse of dimensionality'."*

We may not be accustomed to Bellman's way of thinking. Generally, we tend to believe that more data is better; for example, the more independent variables that we have, the easier or "better" our solution will be. Practical experience in machine learning and data analysis may suggest otherwise.

We can explore Bellman's summary in two ways: by using a simple geometrical object, and by using probability of occurrence. The first way is a *hypercube* and the second way is a Gaussian, or normal, probability distribution.

36.1 Exploring Dimensionality – Hypercubes

One way to explore the effects of dimensionality is to confine our data to take values only within a restricted range. In one dimension, we can study behavior by confining the data to a line segment. In two dimensions, we can confine the data to a square. In three dimensions, we can confine the data to a cube. In general, this is confining our data to have values that are on or within a *hypercube*.

A hypercube is a graph in N dimensions and a *unit* hypercube is a hypercube that has equal, "standard" lengths of edges. We will examine only the vertices in a Cartesian coordinate frame. Of the several alternative definitions of the coordinates of the vertices, and consequently of the length of each edge, we will use:

- A hypercube vertex $\vec{v} \in \mathbb{R}^n$ has coordinates $v_k = \pm 1$
- Vertex $\vec{v}_i$ is adjacent to vertex $\vec{v}_{j \neq i}$ if and only if $\vec{v}_i$ and $\vec{v}_{j \neq i}$ differ in exactly one entry, or

$$\vec{v}_i \cdot \vec{v}_{j \neq i} = n - 2$$

- An edge that is adjacent to $\vec{v}_i$ and $\vec{v}_{j \neq i}$ has length 2

A hypercube in $n = 0$ is a point. Figure 36.1 shows hypercubes for $n = 1, 2, 3$. These are simple, familiar geometrical objects.

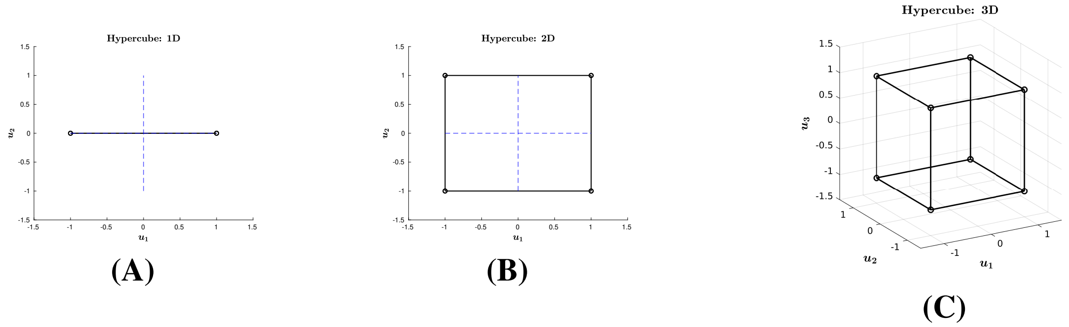


Figure 36.1: Unit hypercubes. (A) Dimension 1 is a line segment. (B) Dimension 2 is a square. (C) Dimension 3 is a cube.

36.2 Exploring Dimensionality – Hypercube Labels

The number of vertices for a hypercube in n dimensions is easily calculated. For each vertex $\vec{v}$, the first entry $v_1 = \pm 1$; it can have 2 values. Likewise, vertex $v_2 = \pm 1$, and so on to $v_n = \pm 1$. The number of vertices, or the cardinality of the set of vertices $\mathcal{V}$, is

$$|\mathcal{V}| = 2 \times 2 \times \cdots \times 2 = 2^n \tag{36.1}$$

We can deduce, from Equation 36.1, that the hypercube number of vertices of a hypercube grows exponentially with the number of dimensions.

We can consider are related problem: if we have n binary variables, how many ways are there to assign values to the variables? The answer comes from Equation 36.1, which is that there are 2^n binary-variable assignments.

This is one result of the *curse of dimensionality*: the number of vertices of a hypercube, and the number of ways to assign binary variables, grows exponentially with the dimensionality of the variables.

Next, we can ask how many ways there are to assign binary values to the vertices of a hypercube. We can use the concept in Equation 36.1 to understand that each vertex can be independently assigned a value, so the answer is

$$2 \times 2 \times \cdots \times 2 = 2^{|\mathcal{V}|} = 2^{2^n} \tag{36.2}$$

From Equation 36.2, we can deduce that the number of ways that binary labels can be assigned to the vertices of a hypercube is a super-exponential function of the dimensionality of the variables.

As a practical matter, suppose that we are trying to train a machine-learning algorithm that has n internal variables, or degrees of freedom. One guide to training, in machine learning, is to use 10 to 20 data observations for each variable. Using a convenient binary notation, the guide could be stated as: use $n \times 2^4$ observations.

Suppose that each variable can be assigned one of two values – such as either 0 or 1 – then there are 2^n internal states in the algorithm. The ratio of the guideline-number of observations and the number of internal states is

$$\frac{n \times 2^4}{2^n} = \frac{n}{2^{n-4}} \tag{36.3}$$

If we have 4 binary variables to train, then the guideline will over-determine the variables by the ratio

$$\frac{4}{2^0} = 4$$

If we have 16 binary variables to train, then the guideline will *under-determine* the variables by the ratio

$$\frac{16}{2^{16-4}} = \frac{1}{256}$$

The guideline may be appropriate for solutions that use linear regression. It is unclear that the guidelines are appropriate for solutions that have even a modest number of internal variables that must have the values learned.

36.3 Exploring Dimensionality – Uniformly Distributed Data

One commonly encountered model of data is that, for each independent variable, the data are uniformly distributed. Suppose that we want to model such data as being on or inside a hypercube. We might ask "where" such data are. To answer this question, we first need to shift and scale the data to fit our model of a unit hypercube.

Real data may not be distributed within the domain $[-1\,,\,1]$ that are the values of each variable. Suppose that we have m observations, gathered into a data vector $\vec{a}_j \in \mathbb{R}^m$, that are uniformly distributed within the domain $[\min(\vec{a}_j)\,,\,\max(\vec{a}_j)]$. We can transform the vector $\vec{a}_j$ to a uniformly distributed variate $u_j \in [-1\,,\,1]$ by the transformation

$$\vec{u}_j = \frac{\vec{a}_j - \min(\vec{a}_j)}{\max(\vec{a}_j) - \min(\vec{a}_j)} \times 2 - 1 \tag{36.4}$$

An important assumption, for interpreting of Equation 36.4, is that each variable is independent and identically distributed (IID). With this assumption, we can gather the n variates of Equation 36.4 into a uniformly distributed vector $\vec{u} \in \mathbb{R}^n$. Our question, of "where" the data are, can be rephrased as:

What is the size d of a hypercube that contains half of the data?

A hypercube that has the length of each edge equal to d has a volume $d \times d \times \cdots \times \; = d^n$. If this hypercube has one-half of the volume of a unit hypercube, then we can solve for the value of d – which is a function of the number of dimensions n – as

$$\begin{array}{rrcl}
 & d^n & = & \frac{1}{2}2^n \\
\equiv & d^n & = & 2^{n-1} \\
\equiv & d^n & = & 2^{-1}2^n \\
\equiv & e^{n\ln(d)} & = & e^{(n-1)\ln(2)} \\
\equiv & n\ln(d) & = & (n-1)\ln(2) \\
\equiv & \ln(d) & = & (n-1)/n\ln(2) \\
\equiv & d & = & 2^{(1-1/n)}
\end{array} \tag{36.5}$$

We can see, from Equation 36.5, that the size d asymptotically approaches the length of the edges of a unit hypercube as the number n of dimensions increases. Another way to think of "where" the data are is to find the thickness T of the "shell" between the unit hypercube and the hypercube that contains one-half of the data. This thickness is the length of a unit hypercube, minus the size d, divided by 2 because the "shell" surrounds the internal hypercube; that is,

$$T = \frac{2-d}{2} \tag{36.6}$$

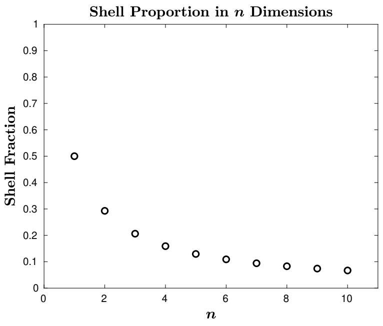


Figure 36.2: Thickness of the shell between a unit hypercube and a hypercube of one-half of the unit hypercube volume. The thickness of the wall decreases exponentially with the dimension of the vector space of the hypercube. For example, a 3D half-volume hypercube has a shell thickness of ≈ 0.2, which implies that one-half of uniformly distributed data in a hypercube are concentrated in a shell that is $\approx 1/10$ of the hypercube diameter.

We can plot the "shell" thickness T of Equation 36.6 as a function of the number n of variables, as shown in Figure 36.2. We see that the thickness exponentially approaches zero. This is a second version of the *curse of dimensionality*: for uniformly distributed data, half of the data are concentrated near the bounding hyperplanes of a hypercube, and most of the central portion has almost none of the data.

36.4 Exploring Dimensionality – Normally Distributed Data

Another commonly encountered model of data is that, for each independent variable, the data are normally distributed. We might ask "where" such data are. An important assumption is that each variable is independent and identically distributed (IID), so that we can apply ordinary statistics to our question.

Let us suppose that each observation has variates u_j that each have a Gaussian distribution, so that for each variate we have a zero mean, written as $\mu = 0$, and a unit variance, written as $\sigma = 1$. This is the distribution, for example, after we standardize our data. After we gather the variates into the multivariate random variable $\vec{u}$, which has a mean $\vec{0}$ and a unit covariance matrix $\Sigma = I$, we can interpret each observation as an instance of $\vec{u}^T$.

The probability density function of a multivariate normal distribution, with the above assumptions, is parameterized by the dimensionality n and has a vector argument $\vec{u}$; we will write this function as

$$P_n(\vec{u}) \overset{\text{def}}{=} \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{\|\vec{u}\|^2}{2}} \tag{36.7}$$

The maximum density of Equation 36.7 occurs at the argument $\vec{0}$. If we plot this maximum density for discrete values of n, as shown in Figure 36.3, we can see that the maximum density rapidly approaches zero.

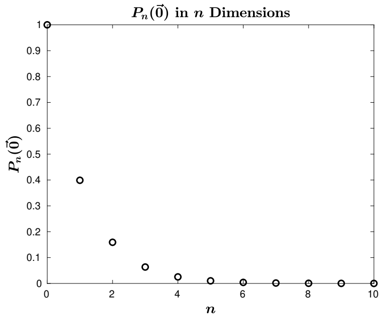


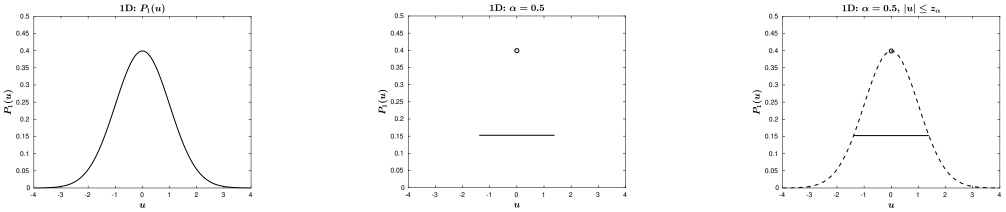
Figure 36.3: For a Gaussian distribution with $\mu = 0$ and $\sigma = 1$, the probability of occurrence $P_n(\vec{0})$ of the zero vector, as a function of the dimension of the vector space of the distribution. The probability $P_n(\vec{0})$ is the "height" of the Gaussian distribution, in n dimensions, that integrates to 1.

This plot suggests that the probability of "finding" standardized data near the origin of $\mathbb{R}^n$ is vanishingly small. What it the probability of finding data that are at some proportion of this maximum density? If we use a hyper-parameter $0 < \alpha < 1$ as the ratio, we can state the problem as that of finding values of a multivariate random variable that are within a density of α of the maximum density. The relevant ratio, using the identity $e^0 = 1$ and the abbreviation z_α, can be written as

$$\begin{aligned}
z_\alpha &\overset{\text{def}}{=} -2\ln(\alpha) \\
\frac{P_n(\vec{u})}{P_n(\vec{0})} &\geq \alpha \\
e^{-\frac{\|\vec{u}\|^2}{2}} &\geq \alpha \\
-\frac{\|\vec{u}\|^2}{2} &\geq \ln(\alpha) \\
\|\vec{u}\|^2 &\leq -2\ln(\alpha) \\
\|\vec{u}\|^2 &\leq z_\alpha \qquad (36.8)
\end{aligned}$$

The value $z_{\alpha=0.5}$ in Equation 36.8 is sometimes called the full-width half maximum (FWHM) value for the Gaussian distribution. It is a practical measure for empirical data and we can use the FWHM to understand effects of dimensionality.

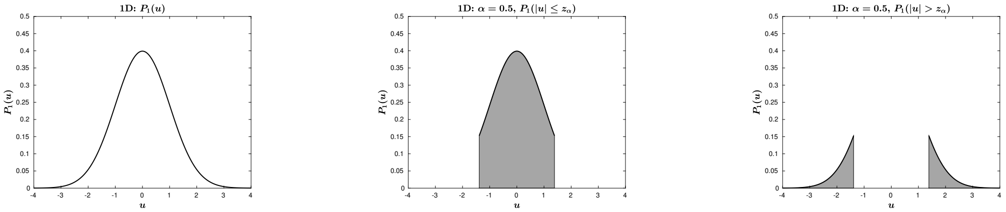
First, let us explore $n = 1$ or the case of a single variate. In Figure 36.4 we can see plots of the probability distribution, plus the maximum and FWHM for $z_{\alpha=0.5}$ both individually and superimposed. These may be familar from prerequisite material.



(A) **(B)** **(C)**

Figure 36.4: The full width half maximum of a Gaussian distribution in 1D. (A) Probability density function $P_1(u)$. (B) The "height" is $P_1(0)$ and the "width" is the interval $[-z_\alpha\,,\, z_\alpha]$ for $\alpha = 1/2$. (C) The "height" and "width" superimposed on the probability density function.

For a univariate distribution, we can also find the cumulative density for a value of the variate u. After numerically integrating $P_1(u)$ between $-z_{\alpha=0.5}$ and $-z_{\alpha=0.5}$, we find that approximately 76% of the probability density occurs within the FWHM domain, which is within the "central" region, as shown in Figure 36.5. Consequently, approximately 24% of the probability density occurs outside the FWHM domain, which is in the "tails" of the probability distribution.



(A) **(B)** **(C)**

Figure 36.5: The cumulative density of a Gaussian distribution in 1D. (A) Probability density function $P_1(u)$. (B) Approximately 76% of the probability of occurrence is within the full-width half maximum, or "central" region, of the Gaussian distribution. (C) Approximately 24% of the probability of occurrence is in the "tails" of the Gaussian distribution.

Next, let us consider a bivariate random variable that has the probability density function $P_2(\vec{u})$. Using the same methods as for a univariate random variable, computations and Figure 36.6 show a major difference from the univariate solution. Here, approximately 50% of the probability density is "central" and within $z_{\alpha=0.5}$; approximately 50% of the probability density is in the "tails" of the distribution.

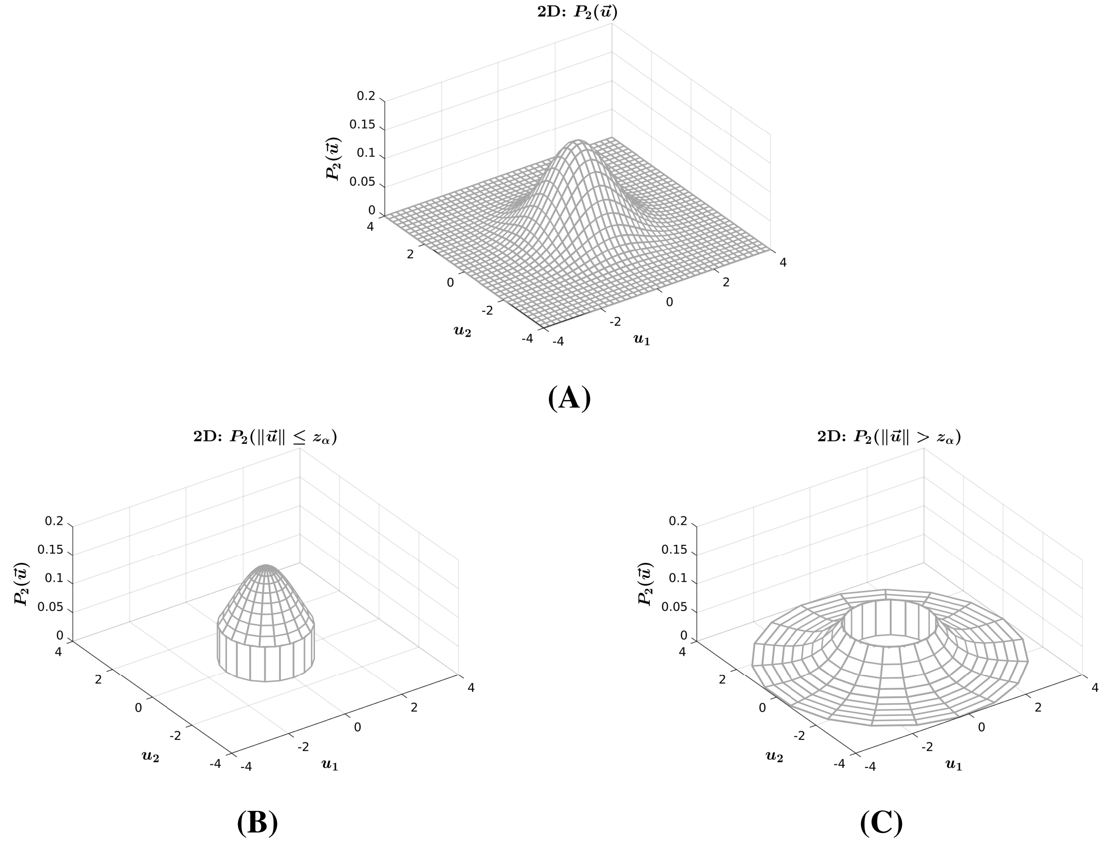
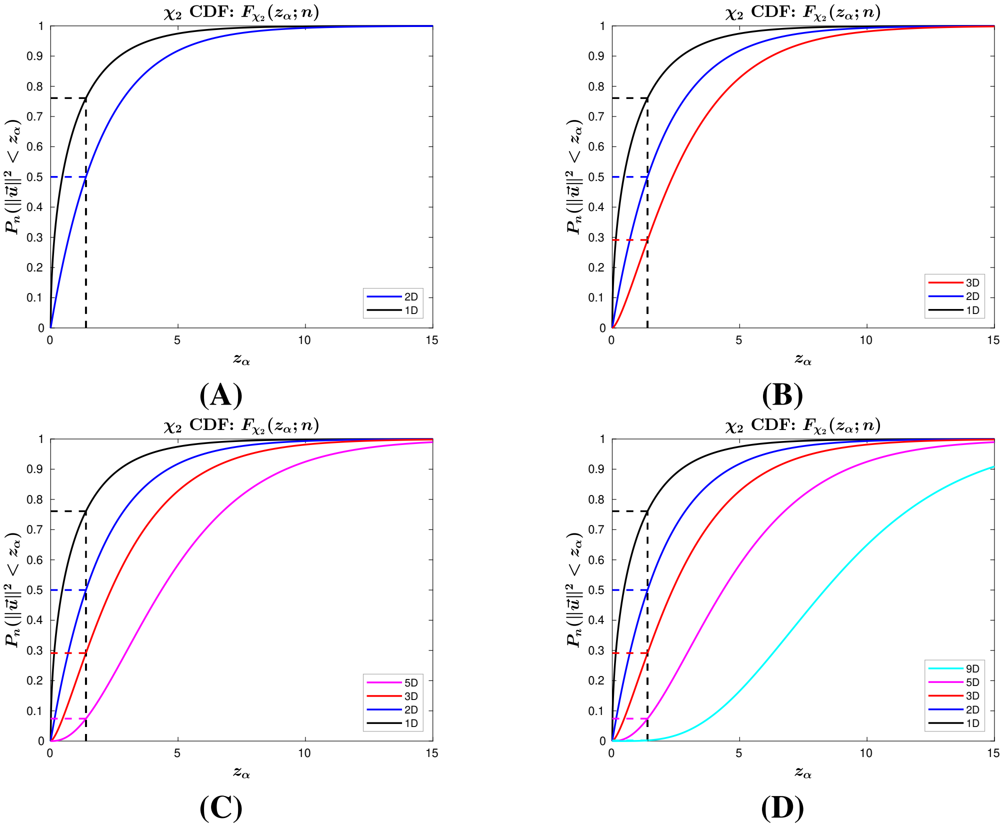


Figure 36.6: The cumulative density function of a Gaussian distribution in 2D. (A) Probability density function $P_2(\vec{u})$. (B) Approximately 50% of the probability of occurrence is within the full-width half maximum, or "central" region, of the Gaussian distribution. (C) Approximately 50% of the probability of occurrence is in the "tails" of the Gaussian distribution.

For $n > 2$, we can use a property of the squared norm $\|\vec{u}\|^2$, which can also be written as

$$z = \|\vec{u}\|^2 = \sum_{j=1}^{n} (u_j)^2 \tag{36.9}$$

IID variates with Gaussian distributions produce a sum z, in Equation 36.9, that follows a χ^2 distribution. The probability that a multivariate $\vec{u}$ is within $z_{\alpha=0.5}$ of the origin is the value cumulative distribution function $F_{\chi_2}(z; n)$. We can plot numerical estimates of the χ^2 cumulative distribution function and note the values at $\alpha = 0.4$ for each plot. Example choices of the dimensionality n, and $z_{\alpha=0.5}$, are shown in Figure 36.7.



(A)

(B)

(C)

(D)

Figure 36.7: The cumulative density function (CDF) of a Gaussian distribution in n dimensions, marking the cumulative density at z_α for $\alpha = 1/2$. The probability of a vector occurring within the full-width half maximum will rapidly decrease with the number n. (A) CDF plots for $n = 1, 2$. (B) CDF plots for $n = 1, 2, 3$. (C) CDF plots for $n = 1, 2, 3, 5$. (D) CDF plots for $n = 1, 2, 3, 5, 9$.

Figure 36.7 shows us that, for normally distributed data, most of the data are in the "tails" of the data and a vanishingly small proportion of the data are "near" the origin. This is a second version of the *curse of dimensionality*: for uniformly distributed data, most of the data "live" in the tails and are statistically less likely to occur near the origin.

36.5 Exploring Dimensionality – Summary

We have explored three ways in which high dimensionality presents challenges in machine learning and data analysis. The first “curse of dimensionality” is in assigning binary values to variables; the number of ways of performing the assignment grows exponentially with the number of variables. The second “curse” is for data that are uniformly distributed in a finite domain; such data are statistically more likely to occur near the limits of each variable than near the mean of the data. The third “curse” is for data that have a normal distribution, also called a Gaussian distribution; Such data are statistically more likely to occur in the “tails” of the data than near the mean.