extend this approach to deal with input spaces having several variables. If we have D input variables, then a general polynomial with coefficients up to order 3 would take the form

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij} x_i x_j + \sum_{i=1}^{D}\sum_{j=1}^{D}\sum_{k=1}^{D} w_{ijk} x_i x_j x_k. \quad (1.74)$$

As D increases, so the number of independent coefficients (not all of the coefficients are independent due to interchange symmetries amongst the x variables) grows proportionally to D^3. In practice, to capture complex dependencies in the data, we may need to use a higher-order polynomial. For a polynomial of order M, the growth in the number of coefficients is like D^M. Although this is now a power law growth, rather than an exponential growth, it still points to the method becoming rapidly unwieldy and of limited practical utility.

Exercise 1.16

Our geometrical intuitions, formed through a life spent in a space of three dimensions, can fail badly when we consider spaces of higher dimensionality. As a simple example, consider a sphere of radius $r = 1$ in a space of D dimensions, and ask what is the fraction of the volume of the sphere that lies between radius $r = 1-\epsilon$ and $r = 1$. We can evaluate this fraction by noting that the volume of a sphere of radius r in D dimensions must scale as r^D, and so we write

$$V_D(r) = K_D r^D \quad (1.75)$$

Exercise 1.18

where the constant K_D depends only on D. Thus the required fraction is given by

$$\frac{V_D(1) - V_D(1-\epsilon)}{V_D(1)} = 1 - (1-\epsilon)^D \quad (1.76)$$

which is plotted as a function of ϵ for various values of D in Figure 1.22. We see that, for large D, this fraction tends to 1 even for small values of ϵ. Thus, in spaces of high dimensionality, most of the volume of a sphere is concentrated in a thin shell near the surface!

As a further example, of direct relevance to pattern recognition, consider the behaviour of a Gaussian distribution in a high-dimensional space. If we transform from Cartesian to polar coordinates, and then integrate out the directional variables, we obtain an expression for the density $p(r)$ as a function of radius r from the origin. Thus $p(r)\delta r$ is the probability mass inside a thin shell of thickness δr located at radius r. This distribution is plotted, for various values of D, in Figure 1.23, and we see that for large D the probability mass of the Gaussian is concentrated in a thin shell.

Exercise 1.20

The severe difficulty that can arise in spaces of many dimensions is sometimes called the *curse of dimensionality* (Bellman, 1961). In this book, we shall make extensive use of illustrative examples involving input spaces of one or two dimensions, because this makes it particularly easy to illustrate the techniques graphically. The reader should be warned, however, that not all intuitions developed in spaces of low dimensionality will generalize to spaces of many dimensions.

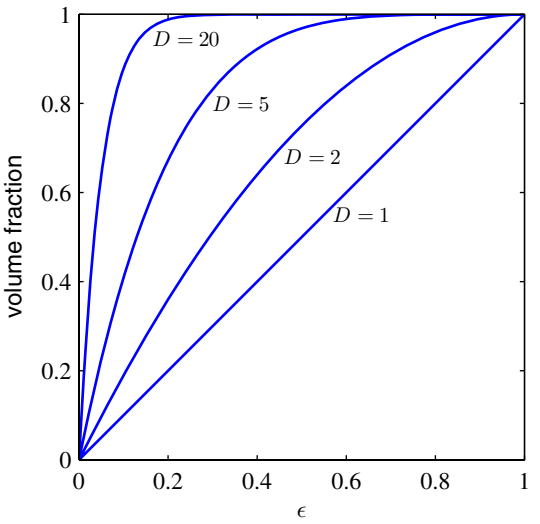

Figure 1.22 Plot of the fraction of the volume of a sphere lying in the range $r = 1-\epsilon$ to $r = 1$ for various values of the dimensionality D.

Although the curse of dimensionality certainly raises important issues for pattern recognition applications, it does not prevent us from finding effective techniques applicable to high-dimensional spaces. The reasons for this are twofold. First, real data will often be confined to a region of the space having lower effective dimensionality, and in particular the directions over which important variations in the target variables occur may be so confined. Second, real data will typically exhibit some smoothness properties (at least locally) so that for the most part small changes in the input variables will produce small changes in the target variables, and so we can exploit local interpolation-like techniques to allow us to make predictions of the target variables for new values of the input variables. Successful pattern recognition techniques exploit one or both of these properties. Consider, for example, an application in manufacturing in which images are captured of identical planar objects on a conveyor belt, in which the goal is to determine their orientation. Each image is a point

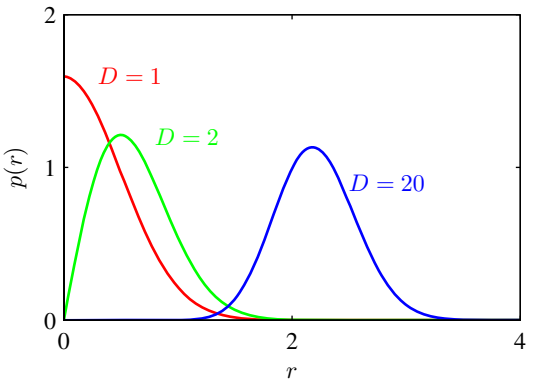

Figure 1.23 Plot of the probability density with respect to radius r of a Gaussian distribution for various values of the dimensionality D. In a high-dimensional space, most of the probability mass of a Gaussian is located within a thin shell at a specific radius.

1.20 (⋆⋆) **www** In this exercise, we explore the behaviour of the Gaussian distribution in high-dimensional spaces. Consider a Gaussian distribution in D dimensions given by

$$p(\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2}\right). \tag{1.147}$$

We wish to find the density with respect to radius in polar coordinates in which the direction variables have been integrated out. To do this, show that the integral of the probability density over a thin shell of radius r and thickness ϵ, where $\epsilon \ll 1$, is given by $p(r)\epsilon$ where

$$p(r) = \frac{S_D r^{D-1}}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{r^2}{2\sigma^2}\right) \tag{1.148}$$

where S_D is the surface area of a unit sphere in D dimensions. Show that the function $p(r)$ has a single stationary point located, for large D, at $\widehat{r} \simeq \sqrt{D}\sigma$. By considering $p(\widehat{r} + \epsilon)$ where $\epsilon \ll \widehat{r}$, show that for large D,

$$p(\widehat{r} + \epsilon) = p(\widehat{r}) \exp\left(-\frac{3\epsilon^2}{2\sigma^2}\right) \tag{1.149}$$

which shows that $\widehat{r}$ is a maximum of the radial probability density and also that $p(r)$ decays exponentially away from its maximum at $\widehat{r}$ with length scale σ. We have already seen that $\sigma \ll \widehat{r}$ for large D, and so we see that most of the probability mass is concentrated in a thin shell at large radius. Finally, show that the probability density $p(\mathbf{x})$ is larger at the origin than at the radius $\widehat{r}$ by a factor of $\exp(D/2)$. We therefore see that most of the probability mass in a high-dimensional Gaussian distribution is located at a different radius from the region of high probability density. This property of distributions in spaces of high dimensionality will have important consequences when we consider Bayesian inference of model parameters in later chapters.

1.21 (⋆⋆) Consider two nonnegative numbers a and b, and show that, if $a \leqslant b$, then $a \leqslant (ab)^{1/2}$. Use this result to show that, if the decision regions of a two-class classification problem are chosen to minimize the probability of misclassification, this probability will satisfy

$$p(\text{mistake}) \leqslant \int \{p(\mathbf{x}, \mathcal{C}_1)p(\mathbf{x}, \mathcal{C}_2)\}^{1/2} \, \mathrm{d}\mathbf{x}. \tag{1.150}$$

1.22 (⋆) **www** Given a loss matrix with elements L_{kj}, the expected risk is minimized if, for each $\mathbf{x}$, we choose the class that minimizes (1.81). Verify that, when the loss matrix is given by $L_{kj} = 1 - I_{kj}$, where I_{kj} are the elements of the identity matrix, this reduces to the criterion of choosing the class having the largest posterior probability. What is the interpretation of this form of loss matrix?

1.23 (⋆) Derive the criterion for minimizing the expected loss when there is a general loss matrix and general prior probabilities for the classes.

reasoning, the density of ρ at distance r is proportional to r^{d-1} in d dimensions. Solving $\int_{r=0}^{r=1} cr^{d-1}dr = 1$ (the integral of density must equal 1) we should set $c = d$. Another way to see this formally is that the volume of the radius r ball in d dimensions is $r^d V_d$, where V_d is the volume of the unit ball. The density at radius r is exactly $\frac{d}{dr}(r^d V_d) = dr^{d-1}V_d$. So, pick $\rho(r)$ with density equal to dr^{d-1} for r over $[0, 1]$.

We have succeeded in generating a point

$$\mathbf{y} = \rho\frac{\mathbf{x}}{|\mathbf{x}|}$$

uniformly at random from the unit ball S by using the convenient spherical Gaussian distribution. In the next sections, we will analyze the spherical Gaussian in more detail.

2.6 Gaussians in High Dimension

A 1-dimensional Gaussian has its mass close to the origin. However, as the dimension is increased something different happens. The d-dimensional spherical Gaussian with zero mean and variance σ^2 in each coordinate has density function

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\,\sigma^d}\exp\left(-\frac{|\mathbf{x}|^2}{2\sigma^2}\right).$$

The value of the density is maximum at the origin, but there is very little volume there. When $\sigma = 1$, integrating the probability density over a unit ball centered at the origin yields nearly zero mass since the volume of such a ball is negligible. In fact, one needs to increase the radius of the ball to nearly $\sqrt{d}$ before there is a significant nonzero volume and hence significant probability mass. If one increases the radius much beyond $\sqrt{d}$, the integral barely increases even though the volume increases since the probability density is dropping off at a much higher rate. The following theorem states this formally that nearly all the probability is concentrated in a thin annulus of width $O(1)$ at radius $\sqrt{d}$.

Theorem 2.8 (Gaussian Annulus Theorem) *For a d-dimensional spherical Gaussian with unit variance in each direction, for any $\beta \leq \sqrt{d}$, all but at most $3e^{-c\beta^2}$ of the probability mass lies within the annulus $\sqrt{d} - \beta \leq |\mathbf{x}| \leq \sqrt{d} + \beta$, where c is a fixed positive constant.*

For a high-level intuition, note that $E(|\mathbf{x}|^2) = \sum_{i=1}^{d} E(x_i^2) = dE(x_1^2) = d$, so the mean squared distance of a point from the center is d. The Gaussian Annulus Theorem says that the points are tightly concentrated. We call the square root of the mean squared distance, namely $\sqrt{d}$, the radius of the Gaussian.

To prove the Gaussian Annulus Theorem we make use of a tail inequality for sums of independent random variables of bounded moments.

1.19 The volume of the cube is $(2a)^D$. Combining this with (1.143) and (1.144) we obtain (1.145). Using Stirling's formula (1.146) in (1.145) the ratio becomes, for large D,

$$\frac{\text{volume of sphere}}{\text{volume of cube}} = \left(\frac{\pi e}{2D}\right)^{D/2} \frac{1}{D} \tag{44}$$

which goes to 0 as $D \to \infty$. The distance from the center of the cube to the mid point of one of the sides is a, since this is where it makes contact with the sphere. Similarly the distance to one of the corners is $a\sqrt{D}$ from Pythagoras' theorem. Thus the ratio is $\sqrt{D}$.

1.20 Since $p(\mathbf{x})$ is radially symmetric it will be roughly constant over the shell of radius r and thickness ϵ. This shell has volume $S_D r^{D-1}\epsilon$ and since $\|\mathbf{x}\|^2 = r^2$ we have

$$\int_{\text{shell}} p(\mathbf{x})\,\mathrm{d}\mathbf{x} \simeq p(r) S_D r^{D-1}\epsilon \tag{45}$$

from which we obtain (1.148). We can find the stationary points of $p(r)$ by differentiation

$$\frac{\mathrm{d}}{\mathrm{d}r}p(r) \propto \left[(D-1)r^{D-2} + r^{D-1}\left(-\frac{r}{\sigma^2}\right)\right] \exp\left(-\frac{r^2}{2\sigma^2}\right) = 0. \tag{46}$$

Solving for r, and using $D \gg 1$, we obtain $\widehat{r} \simeq \sqrt{D}\sigma$.

Next we note that

$$\begin{aligned} p(\widehat{r}+\epsilon) &\propto (\widehat{r}+\epsilon)^{D-1} \exp\left[-\frac{(\widehat{r}+\epsilon)^2}{2\sigma^2}\right] \\ &= \exp\left[-\frac{(\widehat{r}+\epsilon)^2}{2\sigma^2} + (D-1)\ln(\widehat{r}+\epsilon)\right]. \end{aligned} \tag{47}$$

We now expand $p(r)$ around the point $\widehat{r}$. Since this is a stationary point of $p(r)$ we must keep terms up to second order. Making use of the expansion $\ln(1+x) = x - x^2/2 + O(x^3)$, together with $D \gg 1$, we obtain (1.149).

Finally, from (1.147) we see that the probability density at the origin is given by

$$p(\mathbf{x} = \mathbf{0}) = \frac{1}{(2\pi\sigma^2)^{1/2}}$$

while the density at $\|\mathbf{x}\| = \widehat{r}$ is given from (1.147) by

$$p(\|\mathbf{x}\| = \widehat{r}) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{\widehat{r}^2}{2\sigma^2}\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{D}{2}\right)$$

where we have used $\widehat{r} \simeq \sqrt{D}\sigma$. Thus the ratio of densities is given by $\exp(D/2)$.

Since the variance σ is the same for all classes, the SI decision can be equally based on the Euclidean distance, e.g.,

$$s_e = ||\boldsymbol{x} - \boldsymbol{\mu}_k||^2,$$

where we use s_e to denote the score based on the Euclidean distance.

Cosine approximation

We will show that in a high-dimensional space, the Euclidean distance is well approximated by the cosine distance, under the linear Gaussian assumption.

First notice that the Gaussian annulus theorem [4] states that for a d-dimensional Gaussian distribution with the same variance ϵ in each direction, nearly all the probability mass is concentrated in a thin annulus of width $O(1)$ at radius $\sqrt{\epsilon d}$, as shown in Figure 1. This slightly anti-intuitive result indicates that in a high-dimensional space, most of the samples from a Gaussian tend to be in the same length. Rigid proof for this theorem can be found in [4].

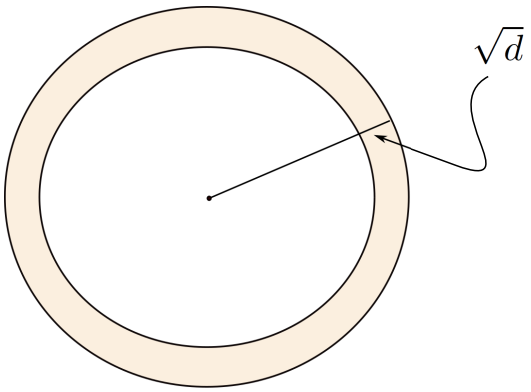

Figure 1: Gaussian annulus theorem [4]: for a d-dimensional multi-variant Gaussian with unit variance in all directions, for any $\beta \leq \sqrt{d}$, all but at most $3e^{-c\beta^2}$ of the probability mass lies within the annulus $\sqrt{d} - \beta \leq ||x|| \leq \sqrt{d} + \beta$, where c is a fixed positive constant. The color region shown in the figure represents the annulus.

Now we rewrite the Euclidean score as follows:

$$s_e = ||\boldsymbol{x}||^2 + ||\boldsymbol{\mu}_k||^2 - 2\cos(\boldsymbol{x}, \boldsymbol{\mu}_k)||\boldsymbol{x}||\ ||\boldsymbol{\mu}_k||,$$

since $||\boldsymbol{\mu}_k|| \approx \sqrt{\epsilon d}$, $\cos(\boldsymbol{x}, \boldsymbol{\mu}_k)$ will be the only term that discriminates the probability that $\boldsymbol{x}$ belongs to different class k. This leads to the cosine score:

$$s_c = \cos(\boldsymbol{x}, \boldsymbol{\mu}_k).$$

This result provides the rationality of the cosine score. It should be noted that this approximation is only valid for high-dimensional data, and the class means must be from a Gaussian with a zero mean. Therefore, data centralization is important for cosine scoring.

2.2 Normalized likelihood (NL) for SV

For SV tasks, our goal is to test two outcomes and check which one is more probable: $\{$ H_0: $\boldsymbol{x}$ belongs to class k; H_1: $\boldsymbol{x}$ belongs to any class other than k $\}$. Again following the MAP principle, the optimal decision can be derived from the posterior $p(H_0|\boldsymbol{x})$. Assuming an equal prior, this leads to:

$$p(H_0|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|H_0)}{p(\boldsymbol{x}|H_0) + p(\boldsymbol{x}|H_1)}.$$