Name: Hall Liu
Date: April 4, 2014

1

Over such a thin shell of thickness ϵ, we can assume the density function to be roughly constant. Then, evaluating the density function at any point x with norm r gives

$$\frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(\frac{-1}{2} x^T \sigma^{-2} I x\right) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(\frac{-r^2}{2\sigma^2}\right)$$

The hypervolume of the surface of the sphere is $S_d r^{d-1}$ (since it's a $d-1$-dimensional set), so the volume of the shell is approximately $\epsilon S_d r^{d-1}$. The integral of the density over the shell is approximately the volume of the shell times the value of the density, or

$$\epsilon \frac{S_d r^{d-1}}{(2\pi\sigma^2)^{d/2}} \exp\left(\frac{-r^2}{2\sigma^2}\right)$$

To find the maximum of f, differentiate it by r to obtain

$$\frac{(d-1)S_d r^{d-2}}{(2\pi\sigma^2)^{d/2}} \exp\left(\frac{-r^2}{2\sigma^2}\right) - \frac{S_d r^d}{\sigma^2(2\pi\sigma^2)^{d/2}} \exp\left(\frac{-r^2}{2\sigma^2}\right)$$

Setting this equal to 0 and canceling gives

$$0 = (d-1) - \frac{r^2}{\sigma^2} \implies r = \pm\sigma\sqrt{d-1}$$

Now, to check that this is a maximum, taking the second derivative gives us a common, positive multiplier times the following:

$$(d-1)(d-2) - \frac{r^2}{\sigma^2}(2d-1) + \frac{r^4}{\sigma^4}$$

Plugging in the previously obtained value for the stationary point, we have

$$d^2 - 3d + 2 - (d-1)(2d-1) + (d-1)^2 = -2d + 2$$

which becomes negative for large d. Thus, the point is a maximum, and we can estimate it by $\sigma\sqrt{d}$ for large d.

0.1 2

By the AM-GM inequality, we have $\frac{a+b}{2} \leq \sqrt{ab}$. Since the minimum of a and b is bounded above by the arithmetic mean, we thus also have $\min(a,b) \leq \sqrt{ab}$.

Now, we have that $P(\text{error}) = P(\text{error}|Y=1)P(Y=1) + P(\text{error}|Y=2)P(Y=2)$. If we let Σ_i be the set which the classifier classifies as $\widehat{Y} = i$, then the above is equal to

$$\int_{\Sigma_2} P(x|Y=1)P(Y=1)dx + \int_{\Sigma_1} P(x|Y=2)P(Y=2)dx$$

Over Σ_1, we have that $P(x|Y=1)P(Y=1) \leq P(x|Y=2)P(Y=2)$ by the form of the Bayes classifier, and vice versa for Σ_2. Thus, the integrals above are bounded above by

$$\begin{aligned}
&\frac{1}{2}\int_{\Sigma_2} P(x|Y=1)P(Y=1) + P(x|Y=2)P(Y=2)dx + \frac{1}{2}\int_{\Sigma_1} P(x|Y=1)P(Y=1) + P(x|Y=2)P(Y=2)dx \\
&= \frac{1}{2}\int_{\mathbb{R}^n} P(x|Y=1)P(Y=1) + P(x|Y=2)P(Y=2)dx
\end{aligned}$$

if we replace the integrands with the average of the two integrands. Now, using the AM-GM inequality, we have that this is bounded above by

$$\begin{aligned}\int_{\mathbb{R}^n} \sqrt{P(x|Y=1)P(Y=1)P(x|Y=2)P(Y=2)}dx &= \sqrt{P(Y=1)P(Y=2)} \int_{\mathbb{R}^n} \sqrt{P(x|Y=1)P(x|Y=2)}dx \\ &\leq \frac{1}{2} \int_{\mathbb{R}^n} \sqrt{P(x|Y=1)P(x|Y=2)}dx\end{aligned}$$

by another application of AM-GM to $\sqrt{P(Y=1)P(Y=2)}$, noting that the average of $P(Y=1)$ and $P(Y=2)$ is $\frac{1}{2}$.

3

a

The decision boundary is the point at which $P(x|Y=1)P(Y=1) = P(x|Y=2)P(Y=2)$, or $\pi_1 \exp\left((x-\mu_1)^2/(2\sigma^2)\right) = \pi_2 \exp\left((x-\mu_2)^2/(2\sigma^2)\right)$. Solving for x, we have

$$x^* = \frac{2\sigma \log \frac{\pi_1}{\pi_2} + \mu_2^2 - \mu_1^2}{2(\mu_2 - \mu_1)}$$

b

The error probability is

$$\int_{-\infty}^{x^*} P(x|Y=2)P(Y=2)dx + \int_{x^*}^{\infty} P(x|Y=1)P(Y=1)dx = \pi_2 \Phi\left(\frac{x^* - \mu_2}{\sigma}\right) + \pi_1 - \pi_1 \Phi\left(\frac{x^* - \mu_1}{\sigma}\right)$$

Rewrite $x^* - \mu_2 = \frac{2\sigma \log \frac{\pi_1}{\pi_2}}{2(\mu_2 - \mu_1)} + \frac{\mu_2 + \mu_1}{2} - \mu_2$. Then, dividing this by σ produces a term constant in σ and a term that goes to $-\infty$ as $\sigma \to 0$. Thus, the π_2 term in the error vanishes. Similarly, $\frac{x^* - \mu_1}{\sigma}$ goes to ∞ as $\sigma \to 0$, which means that the overall error term approaches $\pi_1 - \pi_1 = 0$.

c

As $\pi_1 \to 0$, the log term in the decision boundary approaches $-\infty$, dragging the decision boundary with it, as all other things are fixed. In this case, always classifying things as class 2 would produce an error rate of π_1, which is known to be small.

d

Let S_1 be the set of points in class 1 and S_2 the same for class 2. If we predict class 1, the expected loss is $E_X(L_{2,1}P(Y=2|X)$. If we predict class 2, then it is $E_X(L_{1,2}P(Y=1|X))$. If we minimize this pointwise, we should predict class 1 when $L_{2,1}P(Y=2|X) \leq L_{1,2}P(Y=1|X)$, and class 2 otherwise.

Rewriting the conditional probabilities, the prediction is then $\operatorname{argmax}\left(L_{2,1}P(X|Y=2)P(Y=2), L_{1,2}P(X|Y=1)P(Y=1)\right)$. Setting these equal to find the boundary, we have

$$L_{2,1}\pi_2 \exp\left((x-\mu_2)^2/(2\sigma^2)\right) = L_{1,2}\pi_1 \exp\left((x-\mu_1)^2/(2\sigma^2)\right)$$

Solving for x, we obtain

$$x^* = \frac{2\sigma \log \frac{L_{1,2}\pi_1}{L_{2,1}\pi_2} + \mu_2^2 - \mu_1^2}{2(\mu_2 - \mu_1)}$$

Now, if we choose $L_{1,2}$ to be proportional to π_1^{-1} and choose $L_{2,1}$ to be proportional to π_2^{-1} (possibly with different constants), we will avoid the degeneracy problem.

e

Using these values, we have that $x^* = 1$, so plugging into the expression for the error rate from (b) gives $0.5\Phi(-1) + 0.5 - 0.5\Phi(1) = 0.1587$

4

The expected loss of some decision function $\widehat{H}$ is $E_{X,Y}(L(\widehat{H}(x),Y)$. Writing out the integral gives

$$\int_{\mathbb{R}^n}\sum_{k=1}^{K}L(\widehat{H}(x),k)p(x,k)dx=\int_{\mathbb{R}^n}p(x)\sum_{k=1}^{K}L(\widehat{H}(x),k)P(k|x)dx$$

Define R_i as $\{x\in\mathbb{R}^n|\widehat{H}(x)=i\}$. Then, suppose that there exists some subset of nonzero measure A of R_i such that $\sum_{k=1}^{K}L(i,k)P(k|x)\geq\sum_{k=1}^{K}L(j,k)P(k|x)$. Decompose the above integral as

$$\int_{\mathbb{R}^n-A}p(x)\sum_{k=1}^{K}L(\widehat{H}(x),k)P(k|x)dx+\int_{A}p(x)\sum_{k=1}^{K}L(i,k)P(k|x)dx$$

Then, changing the class of A from i to j will result in a strict decrease in the expected loss, which means that the optimal decision rule must satisfy, for each i, $\sum_{k=1}^{K}L(i,k)P(k|x)\geq\sum_{k=1}^{K}L(j,k)P(k|x)$ for all $j\neq i$ and almost all $x\in R_i$. The rule given satisfies this condition, and since there are a finite number of classes, changing the rule on a set of measure zero from the region associated with each class will not result in a change in the expected loss. Thus, the rule given is optimal.

5

a

We have $\mathrm{cov}(X_2,X_2)=v_2$, so writing $X_2=\alpha X_1+Z$ gives

$$\mathrm{cov}(\alpha X_1+Z,\alpha X_1+Z)=\mathrm{cov}(\alpha X_1,\alpha X_1)+0+\mathrm{cov}(Z,Z)=\alpha^2v_1+\mathrm{var}(Z)=v_2$$

In addition, using $\mathrm{cov}(X_1,X_2)=a$ gives us

$$a=\mathrm{cov}(X_1,\alpha X_1+Z)=\alpha v_1+0\implies\alpha=\frac{a}{v_1}$$

Substituting into the first equation gives us $\mathrm{var}(Z)=v_2-\frac{a^2}{v_1}$

Finally, since $E(X_1)=E(X_2)=0$, we must also have $E(Z)=0$.

b

We know that Z is Gaussian because we can express Z as the sum of two Gaussians, $X_2-\alpha X_1$. Thus, it is independent from X_1 because their covariance is zero. The variance was derived above.

c

We have $E(X_2|X_1=x)=E(\alpha x+Z)=\alpha x$ and $\mathrm{var}(X_2|X_1=x)=\mathrm{var}(\alpha x+Z)=\mathrm{var}(Z)$.

d

Since the samples are all independent, the cross terms in the formula for the variance of a sum disappear and we get

$$\mathrm{var}(\widehat{a})=\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(X_{i1}X_{i2})=\frac{1}{n}\mathrm{var}(X_1X_2)$$

Then, we have $\mathrm{var}(X_1X_2)=E((X_1X_2)^2)-E(X_1X_2)^2=E((X_1X_2)^2)-a^2$, since the means are zero.

To obtain the first term, note that it is a cross moment and can be obtained by applying $\frac{\partial^4}{\partial x_1^2\partial x_2^2}$ to the mgf of the distribution and evaluating at zero. The mgf itself is

$$\exp\left(\frac{1}{2}\begin{pmatrix}x_1 & x_2\end{pmatrix}\begin{pmatrix}v_1 & a\\ a & v_2\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix}\right)=\exp\left(\frac{1}{2}(x_1^2v_1+2x_1x_2a+x_2^2v_2)\right)$$

Differentiating this wrt x_1 twice and wrt x_2 once gives

$$ax_1v_1T+x_1v_1(x_1v_1+x_2a)(x_2v_2+x_1a)T+v_1(x_2v_2+x_1a)T$$

where T is the exponential term from above. Then, taking the derivatives and then evaluating gives v_1v_2, so the variance of $\widehat{a}$ is $\frac{1}{n}(v_1v_2-a^2)$.

6