2.1 If $\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)$, then

$$\begin{aligned}\|t_k - \hat{y}\|^2 &= (1 - \hat{y}_k)^2 + \sum_{i \neq k} \hat{y}_i^2 \\ &= 1 - 2\hat{y}_k + \|\hat{y}\|^2\end{aligned}$$

So the k that minimizes this will maximize $\hat{y}_k$.

2.2 The Bayes classifier picks the value of G_k which maximizes $\Pr(G_k|X = x)$. In this case, the boundary consists of points for which $\Pr(\text{orange}|X = x) = \Pr(\text{blue}|X = x) = \frac{1}{2}$. Letting $f_o(x), f_b(x)$ be the density functions for the blue and orange groups, respectively, we find

$$\Pr(\text{orange}|X = x) = \frac{f_o(x)}{f_b(x) + f_o(x)},$$

and so the decision boundary consists of points satisfying $f_o(x) = f_b(x)$.

2.3 Let $f(x)$ be the density function of the distance of a single point selected uniformly from the unit p-ball and $F(x)$ the cdf. If $x_1, \dots, x_N$ are iid points, let $x_{(c)}$ denote the closest point to the origin. Then

$$\begin{aligned}\Pr(d(0, x_{(c)}) \leq r) &= 1 - \Pr(d(0, x_i) \geq r \forall i) \\ &= 1 - \prod_i \Pr(d(0, x_i) \geq r) \\ &= 1 - \prod_i (1 - F(r)) \\ &= 1 - (1 - F(r))^N\end{aligned}$$

The median of the closest distance is the r for which this is one half, so

$$1 - F(r_{\text{med}}) = \left(\frac{1}{2}\right)^{\frac{1}{N}}.$$

The density function expressed for radial coordinates has the form $\frac{1}{\text{Vol}(D^p)} r^{p-1} g(\theta_1, \dots, \theta_{p-1})$, where the angles parametrize the sphere S^{p-1}. So we find that

$$\begin{aligned}F(r) &= \frac{1}{\text{Vol}(D^p)} \int_0^r \int_{S^{p-1}} \rho^{p-1} g(\overline{\theta}) d\rho d\overline{\theta} \\ &= \frac{1}{\text{Vol}(D^p)} \left(\int_0^r \rho^{p-1} d\rho\right) \left(\int_{S^{p-1}} C g(\overline{\theta}) d\overline{\theta}\right) \\ &= \frac{\text{Vol}(S^{p-1})}{p \text{Vol}(D^p)} r^p\end{aligned}$$

But the form of the volume form shows

$$\mathrm{Vol}(D^p) = \mathrm{Vol}(S^{p-1}) \int_0^1 \rho^{p-1} d\rho = \frac{1}{p}\mathrm{Vol}(S^{p-1}),$$

so

$$F(r) = r^p$$

and hence

$$r_{\text{med}} = \left(1 - \left(\frac{1}{2}\right)^{\frac{1}{N}}\right)^{\frac{1}{p}}.$$

2.4 The density function has the form $Ce^{-\frac{1}{2}\|x\|^2}$, which is clearly invariant under orthonormal basis changes. Hence the projections of points sampled from this distribution onto any unit vector must be $\mathcal{N}(0,1)$. The expected square distance of these z_i is just the variance of $\mathcal{N}(0,1)$, which is 1. The expected square distance of x_0 from the origin is p, as mentioned in the problem.

So if you fix a *particular direction*, most points from a multivariate normal won't lie very far along it.

2.5.a The first equality is the definition. For the second, we expand

$$(y_0 - \hat{y}_0)^2 = y_0^2 - 2y_0\hat{y}_0 + \hat{y}_0^2.$$

Taking $E_{\mathcal{T}}$ and noting that y_0 has no dependence on $\mathcal{T}$, we arrive at

$$\begin{aligned} &y_0^2 - 2y_0 E_{\mathcal{T}}[\hat{y}_0] + E_{\mathcal{T}}[\hat{y}_0^2] \\ =&y_0^2 - 2y_0 E_{\mathcal{T}}[\hat{y}_0] + E_{\mathcal{T}}[\hat{y}_0 - E_{\mathcal{T}}[\hat{y}_0]]^2 + (E_{\mathcal{T}}[\hat{y}_0])^2 \end{aligned}$$

Now take $E_{y_0|x_0}$ to obtain

$$\begin{aligned} &\mathrm{Var}(y_0|x_0) + (x_0^T\beta)^2 - 2x_0^T\beta E_{\mathcal{T}}[\hat{y}_0] + \mathrm{Var}_{\mathcal{T}}(\hat{y}_0) + (E_{\mathcal{T}}[\hat{y}_0])^2 \\ =&\mathrm{Var}(y_0|x_0) + E_{\mathcal{T}}[x_0^T\beta - \hat{y}_0]^2 + \mathrm{Var}_{\mathcal{T}}(\hat{y}_0), \end{aligned}$$

which is the second and third lines.

To obtain the fourth line, we use the fact that ε is $N(0,\sigma^2)$ to write $\mathrm{Var}(y_0|x_0) = \mathrm{Var}(x_0 + \varepsilon|x_0) = \sigma^2$ and the fact that least squares in unbiased here to equate the bias term with 0. This leaves the variance $\mathrm{Var}_{\mathcal{T}}(\hat{y}_0)$. Since there is no bias, the definition gives that

$$\begin{aligned}
\mathrm{Var}_{\mathcal{T}}(\hat{y}_0) &= \mathrm{Var}_{\mathcal{T}}(\sum_i l_i(x_0)\varepsilon_i) \\
&= \sum_{i,j} \mathrm{Cov}_{\mathcal{T}}(l_i(x_0)\varepsilon_i, l_j(x_0)\varepsilon_j) \\
&= \sum_{i,j} E_{\mathcal{T}}[\varepsilon_i\varepsilon_j] E_{\mathcal{T}}[l_i(x_0)l_j(x_0)] \\
&= \sigma^2 \sum_i E_{\mathcal{T}}[l_i(x_0)^2)] \\
&= \sigma^2 E_{\mathcal{T}}\big[\sum_i l_i(x_0)^2\big]
\end{aligned}$$

Now, since $l_i(x_0)$ is the ith row of $X(X^TX)^{-1}x_0$, we find that

$$\sum_i l_i(x_0)^2 = x_0^T(X^TX)^{-T}X^TX(X^TX)^{-1}x_0 = x_0^T(XX^T)^{-1}x_0$$

which gives the final line.

2.5.b Two facts: If A is any $n \times n$ matrix and b is a column vector, we find that

$$b^TAb = \sum_{i,j} A_{ij}b_jb_j$$

. Also that if A, B are symmetric, then

$$\sum_{ij} A_{ij}B_{ij} = \sum_i \sum_j A_{ij}B_{ji} = \sum_i (AB)_{ii} = \mathrm{trace}(AB).$$

So (assuming $E[X] = 0$)

$$\begin{aligned}
E_{x_0}[x_0^T\mathrm{Cov}(X)^{-1}x_0] &= \sum_{i,j} \mathrm{Cov}(X)^{-1}_{ij} E_{x_0}[(x_0)_i(x_0)_j] \\
&= \sum_{i,j} \mathrm{Cov}(X)^{-1}_{ij} \mathrm{Cov}(x_0)_{ij} \\
&= \mathrm{trace}(\mathrm{Cov}(X)^{-1}\mathrm{Cov}(x_0)) \\
&= \mathrm{trace}(I) \\
&= p.
\end{aligned}$$

Plugging this into the equation gives the desired expression.

2.6 Suppose $a_1, \dots, a_k$ is some sample of values. Then if μ is the average of these and $\mathrm{Var}(a)$ is the sample variance

$$\sum_i (a_i - c)^2 = k\,\mathrm{Var}(a) + k(\mu - c)^2$$

Let $\pi_X \mathcal{T}$ denote the set of x values in the training sample. Let $\mathcal{T}_x$ denote the set of y values of training points with x-coordinate x and μ_x be the average of those points. (View $\mathcal{T}_x$ as a multiset if necessary.) Then

$$\begin{aligned} RSS(\theta) &= \sum_i (y_i - f_\theta(x_i))^2 \\ &= \sum_{x \in \pi_X \mathcal{T}} \sum_{y \in \mathcal{T}_x} (y - f_\theta(x))^2 \\ &= \sum_{x \in \pi_X \mathcal{T}} \mathrm{Var}(\mathcal{T}_x) + |\mathcal{T}_x| (\mu_x - f_\theta(x))^2 \\ &= \left(\sum_{x \in \pi_X \mathcal{T}} \mathrm{Var}(\mathcal{T}_x) \right) + WRSS(\theta) \end{aligned}$$

where WRSS is an RSS expression with one sample point for each x value, μ_x, and each x is weighted by $|\mathcal{T}_x|$. Since the first term in the sum has no dependence on θ, minimizing WRSS will minimize RSS.