1. **HTF 2.1**

Assume that $||x||$ is the $L_2$ norm, i.e. $||x|| = \sqrt{\sum_{i=1}^{K} x_i^2}$. Since there are $K$ classes and $t_k$ is an indicator vector, then it has dimension $K$. The individual elements of $t_k$ take on value 1 when $t_{k; k}$ and 0 when $t_{k; l}$ when $l \neq k$.

$$ \arg\min_k ||t_k - \hat{y}|| = \arg\min_k \sqrt{\sum_{l=1}^{K} (t_{k; l} - \hat{y}_l)^2} $$

$$ = \arg\min_k \sum_{l=1}^{K} (t_{k; l} - \hat{y}_l)^2 $$

$$ = \arg\min_k [(1 - \hat{y}_k)^2 + \sum_{l \neq k} \hat{y}_l^2] $$

$$ = \arg\min_k [1 - 2\hat{y}_k + \sum_{l=1}^{K} \hat{y}_l^2] $$

$$ = \arg\min_k [1 - 2\hat{y}_k + ||\hat{y}||^2] = \arg\max_k \hat{y}_k $$

We drop the $||\hat{y}||^2$ term because there is no $k$ dependence.

If we assume that this is an $L_1$ norm, i.e. $||x|| = \sum_{i=1}^{K} |x_i|$. and the elements of $\hat{y}$ are positive and sum to one, then we also get this result.

$$ \arg\min_k ||t_k - \hat{y}|| = \arg\min_k \sum_{l=1}^{K} |t_{k; l} - \hat{y}_l| $$

$$ = \arg\min_k [||1 - \hat{y}_k|| + \sum_{l \neq k} |\hat{y}_l|] $$

$$ = \arg\min_k [(1 - \hat{y}_k) + \sum_{l \neq k} \hat{y}_l] $$

$$ = \arg\min_k 2(1 - \hat{y}_k) = \arg\max_k \hat{y}_k $$

2. **HTF 2.6**

(a) For linear regression, from HTF p. 12,

$$ \hat{f}(x_0) = x_0^T \hat{\beta} = x_0^T (X^T X)^{-1} X^T Y = \sum_{i=1}^{N} x_0^T (X^T X)^{-1} x_i y_i $$

$$ \hat{f}(x_0) = \sum_{x_i \in N_k(x_0)} y_i $$

Thus, $l_i(x_0; X) = x_0^T (X^T X)^{-1} x_i$.

For k-NN,

$$ \hat{f}(x_0) = \frac{1}{k} \sum_{x_i \in N_k(x_0)} y_i $$
where \( N_k(x_0) \) denotes the set of the \( k \) nearest-neighbors of \( x_0 \). Thus, \( l_i(x_0, \mathcal{X}) = 1/k \) if \( x_i \in N_k(x_0) \), and \( l_i(x_0, \mathcal{X}) = 0 \) otherwise.

(b) Decomposing the conditional MSE:
Recall that the conditional bias is \( \left[ f(x_0) - E_{Y|X}[\hat{f}(x_0)] \right] \) and the conditional variance is \( E_{Y|X} \left[ (\hat{f}(x_0) - E_{Y|X}[\hat{f}(x_0)])^2 \right] \). So we can write:

\[
E_{Y|X} \left[ (f(x_0) - \hat{f}(x_0))^2 \right] = E_{Y|X} \left[ (f(x_0) - E_{Y|X}[\hat{f}(x_0)] - (f(x_0) - E_{Y|X}[\hat{f}(x_0)]))^2 \right]
= E_{Y|X} \left[ (f(x_0) - E_{Y|X}[\hat{f}(x_0)])^2 \right] - E_{Y|X} \left[ (f(x_0) - E_{Y|X}[\hat{f}(x_0)])(E_{Y|X}[\hat{f}(x_0)] - \hat{f}(x_0)) \right] + E_{Y|X} \left[ (\hat{f}(x_0) - E_{Y|X}[\hat{f}(x_0)])^2 \right]
\]

But \( E_{Y|X} \left[ (f(x_0) - E_{Y|X}[\hat{f}(x_0)])^2 \right] = (f(x_0) - E_{Y|X}[\hat{f}(x_0)])^2 \) since there is nothing depending on \( Y \) and note that

\[
E_{Y|X} \left[ (f(x_0) - E_{Y|X}[\hat{f}(x_0)])(E_{Y|X}[\hat{f}(x_0)] - \hat{f}(x_0)) \right] = (f(x_0) - E_{Y|X}[\hat{f}(x_0)])(E_{Y|X}[\hat{f}(x_0)] - E_{Y|X}[\hat{f}(x_0)]) = 0
\]

So we are left with \((f(x_0) - E_{Y|X}[\hat{f}(x_0)])^2 \) which is the conditional bias\(^2 + \text{variance}.

3. Draw 10 points from a uniform distribution...

There must be two points on the boundary for points drawn from \([0,1]\). The boundary points are \( \min_i(x_i) \) and \( \max_i(x_i) \).

For \([0,1]^2\), at least 4 points will be on the boundary of the set.

For \([0,1]^3\), at least 6 points will be on the boundary.

For \([0,1]^10\), with probability 1 almost surely, all ten points will be on the boundary of the set.

As the dimensionality gets higher, the probability of being a boundary point increases, very quickly.

Note that in this problem, we assume prior knowledge that the distribution is uniform “rectangular.” If we define the region to be the convex enclosure, then the number of points on the boundary is bigger. For example, in the 3-dimensional case, usually 9-10 points will be on the boundary.

2
4. Suppose events fall into two classes...

(a) As defined on HTF page 21, the Bayes’ Decision Boundary marks where $P(G_1|X = x) = P(G_2|X = x)$.

\[
\begin{align*}
\frac{P(G_1|X = x)}{P(X = x)} &= \frac{P(G_2|X = x)}{P(X = x)} \\
P(X = x|G_1) &= \frac{P(X = x|G_1)}{P(G_2)} \quad \text{(3)}
\end{align*}
\]

The priors $P(G_1)$ and $P(G_2)$ for each of the classes is given, as well as $P(X = x|G_1) \sim \mathcal{N}(u, I)$ and that $P(X = x|G_1) \sim \mathcal{N}(r, I)$. Writing the exponential forms of the Gaussians (see for example, page 86 in HTF) and cancelling, we obtain:

\[
\frac{e^{-(u_1^2 - 2x_1u_1 - 2x_2u_2 + u_2^2)/2}}{e^{-(r_1^2 - 2x_1r_1 - 2x_2r_2 + r_2^2)/2}} = \frac{1}{2} \quad \text{(4)}
\]

Taking the natural log and simplifying, the hyperplane decision boundary has equation:

\[
x_1(u_1 - r_1) + x_2(u_2 - r_2) + \log 2 - (u_1^2 + u_2^2)/2 + (r_1^2 + r_2^2)/2 = 0 \quad \text{(5)}
\]

so we have

\[
\begin{align*}
a &= u_1 - r_1 \\
b &= u_2 - r_2 \\
c &= \log 2 - (u_1^2 + u_2^2)/2 + (r_1^2 + r_2^2)/2
\end{align*}
\]

(b) When the priors are equal, then the right side of equation (4) is equal to one. The decision function is:

\[
x(t(u - r)) > 0.5|u| - |r| = T \quad \Rightarrow \quad \text{Decide 1}
\]

Let $v = x^t(u - r)$, so $P(v|G_1) \sim N(u^t(u - r), |u - r|^2)$ and $P(v|G_2) \sim N(r^t(u - r), |u - r|^2)$ and the fact that linear functions of Gaussians are Gaussian, i.e.

\[
P(v > T) = P\left(\frac{v - \mu_v}{\sigma_v} > \frac{T - \mu_v}{\sigma_v}\right) = Q\left(\frac{T - \mu_v}{\sigma_v}\right)
\]

then the probability of error is:

\[
P_e = P(D_1, G_2) + P(D_2, G_1)
\]

\[
= 0.5 \int_T^\infty p(v|G_2)dv + 0.5 \int_{-\infty}^T p(v|G_1)dv
\]

\[
= 0.5Q\left(\frac{T - r^t(u - r)}{|u - r|}\right) + 0.5(1 - Q\left(\frac{T - u^t(u - r)}{|u - r|}\right))
\]

Plugging in for $T$

\[
\begin{align*}
T - r^t(u - r) &= 0.5|u|^2 - 0.5|r|^2 + ||r||^2 - u^t r \\
&= 0.5||u - r||^2 \\
T - u^t(u - r) &= 0.5|u|^2 - 0.5|r|^2 - ||u||^2 + u^t r \\
&= -0.5||u - r||^2
\end{align*}
\]

3
Noting that for the zero mean Gaussian, \( Q(z) = 1 - Q(-z) \)
\[
P_e = 0.5Q(0.5||u - r||) + 0.5(1 - Q(-0.5||u - r||)) = Q(0.5||u - r||)
\]

5. **This problem skipped for this week. Will be included in a later HW.**

6. **Let G be a rasterized binary image...**

Recall that for a Bernoulli distribution
\[
P(N_i = n) = \epsilon^n(1 - \epsilon)^{(1-n)}
\]

The min probability of error decision rule is the MAP rule:
\[
\arg\max_g \log P(g|v) = \arg\max_g \sum_i \log P(g_i|v_i)
\]

which we simplify using
\[
\log P(g_i|v_i) = \log P(g_i \oplus v_i) \\
= (g_i \oplus v_i) \log \epsilon + (1 - g_i \oplus v_i) \log(1 - \epsilon) \\
= (g_i \oplus v_i) \log(\frac{\epsilon}{1 - \epsilon}) + \log(1 - \epsilon)
\]

Since \( \epsilon < 0.5 \), then \( \frac{\epsilon}{1 - \epsilon} < 1 \) and \( \log(\frac{\epsilon}{1 - \epsilon}) < 0 \)
\[
\arg\max_g \log P(g|v) = \arg\max_g \sum_i (g_i \oplus v_i) \log(\frac{\epsilon}{1 - \epsilon}) = \arg\min_g \sum_i (g_i \oplus v_i)
\]

Note that \( \sum_i (g_i \oplus v_i) \) is exactly the Hamming distance between vectors \( g \) and \( v \). Thus the min prob. of error detection rule corresponds to finding the \( g \) with minimum Hamming distance.

7. **A text containing \( m \) characters...**

So a particular observation of a text of \( m \) characters is given by \( v = [v_1, v_2, \ldots, v_d] \) where \( v_i \) is the count of character \( i \) (note that \( \sum_{i=1}^{d} v_i = m \)). The multinomial distribution has a “vector” parameter so \( \theta = [\theta_{1,1}, \theta_{1,2}, \ldots, \theta_{1,d}] \). This means that:
\[
P(V = v|L = l) = \frac{m!}{v_1!v_2!\ldots v_d!} \theta_{1,1}^{v_1} \theta_{1,2}^{v_2} \cdots \theta_{1,d}^{v_d}
\]

So the MAP rule is to pick the language \( k \) such that
\[
k = \arg\max_L P(v|l)p(l) = \arg\max_L \log P(v|l) + \log p(l)
\]
\[
= \arg\max_L \left[ \log \left( \frac{m!}{v_1!v_2!\ldots v_d!} \right) + \sum_{i=1}^{d} v_i \log \theta_{l,i} + \log p(l) \right]
\]
\[
= \arg\max_L \left[ \log p(l) + \sum_{i=1}^{d} v_i \log \theta_{l,i} \right]
\]
where we drop the factorial terms since there is no dependence on \( l \).
8. Find the ML estimate....
Assume you are given \( n \) data samples.

\[
\hat{\theta} = \arg \max_{\theta} L(\theta) = \arg \max_{\theta} \sum_{i=1}^{n} \log p(x_i|\theta)
\]

\[
= \arg \max_{\theta} \sum_{i=1}^{n} \left[ \log \left( \frac{\theta}{2} \right) - \theta |x_i| \right]
\]

\[
= \arg \max_{\theta} \left[ n \log \left( \frac{\theta}{2} \right) - \theta \sum_{i=1}^{n} |x_i| \right]
\]

Take the derivative with respect to \( \theta \), set to zero, and solve:

\[
\frac{d}{d\theta} L(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} |x_i| = 0
\]

\[
\frac{n}{\theta} = \sum_{i=1}^{n} |x_i|
\]

\[
\hat{\theta} = \frac{n}{\sum_{i=1}^{n} |x_i|}
\]

If you just assumed there was one data point (some people did), that’s ok: the solution is this for \( n = 1 \), i.e. \( \frac{1}{|x_i|} \).