# 1 Ex.2.1

We imagine that we have some input data $x$. Some algorithm assigns to $x$ the probability $y_k$ that $x$ is a member of the $k$-th class. This would explain why we are told to assume that the sum of the $y_k$ is equal to one. (But, if that reason is valid, then we should also have been told that each $y_k \geq 0$.) In fact, neither of these two assumptions is necessary to provide an answer to the question. The hyphen in $K$-classes seems to be a misprint, and should be omitted.

We restate the question, clarifying it, but distancing it even further from its origins. Let $K > 0$ be an integer. For each $k$ with $1 \leq k \leq K$, let $t_k$ be the $K$-dimensional vector that has 1 in the $k$-th position and 0 elsewhere. Then, for any $K$-dimensional vector $y$, the $k$ for which $y_k$ is largest coincides with the $k$ for which $t_k$ is nearest to $y$.

By expanding the quadratic we find that

$$\begin{aligned}
\operatorname{argmin}_k \|y - t_k\| &= \operatorname{argmin}_k \|y - t_k\|^2 \\
&= \operatorname{argmin}_k \sum_{i=1}^{K} \left(y_i - (t_k)_i\right)^2 \\
&= \operatorname{argmin}_k \sum_{i=1}^{K} \left((y_i)^2 - 2y_i\,(t_k)_i + {(t_k)_i}^2\right) \\
&= \operatorname{argmin}_k \sum_{i=1}^{K} \left(-2y_i\,(t_k)_i + {(t_k)_i}^2\right),
\end{aligned}$$

since the sum $\sum_{i=1}^{K} {y_i}^2$ is the same for all classes $k$. Notice that, for each $k$, the sum $\sum_{k=1}^{K} {(t_k)_i}^2 = 1$. Also $\sum y_i\,(t_k)_i = y_k$. This means that

$$\begin{aligned}
\operatorname{argmin}_k \|y - t_k\| &= \operatorname{argmin}_k \left(-2y_k + 1\right) \\
&= \operatorname{argmin}_k \left(-2y_k\right) \\
&= \operatorname{argmax}_k y_k.
\end{aligned}$$

# 2 Ex.2.3

We denote the $N$-tuple of data points by $(x_1, \ldots, x_N)$. Let $r_i = \|x_i\|$. Let $U(A)$ be the set of all $N$-tuples with $A < r_1 < \ldots < r_N < 1$. Ignoring subsets of measure zero, the set of all $N$-tuples is the disjoint union of the $N\,!$ different subsets obtained from $U(0)$ by permuting the indexing set $(1, \ldots, N)$. We will look for $A > 0$ such that the measure of $U(A)$ is half the measure of $U(0)$. The same $A$ will work for each of our $N\,!$ disjoint subsets, and will therefore give the median for the distance of the smallest $x_i$ from the origin.

We want to find $A$ such that

$$\int_{U(A)} dx_1 \ldots dx_N = \frac{1}{2} \int_{U(0)} dx_1 \ldots dx_N$$

We convert to spherical coordinates. Since the coordinate in the unit sphere $S^{p-1}$ contributes the same constant on each side of the equality, obtaining

$$\int_{A<r_1<\ldots<r_N<1} r_1^{p-1} \ldots r_N^{p-1} dr_1 \ldots dr_n = \frac{1}{2} \int_{0<r_1<\ldots<r_N<1} r_1^{p-1} \ldots r_N^{p-1} dr_1 \ldots dr_n.$$

We change coordinates to $s_i = r_i^p$, and the equality becomes

$$\int_{A^p<s_1<\ldots<s_N<1} ds_1 \ldots ds_n = \frac{1}{2} \int_{0<s_1<\ldots<s_N<1} ds_1 \ldots ds_n$$

In the left-hand integral, we change coordinates to

$$t_0 = s_1 - A^p, t_1 = s_2 - s_1, \ldots, t_{N-1} = s_N - s_{N-1}, t_N = 1 - s_N$$

The Jacobian (omitting $t_0$ which is a redundant variable) is a triangular matrix with -1 entries down the diagonal. The absolute value of its determinant, used in the change of variable formula for integration, is therefore equal to 1 .

The region over which we integrate is

$$\sum_{i=0}^{N} t_i = 1 - A^p, \text{ with each } t_i > 0$$

which is an $N$-dimensional simplex scaled down by a factor $(1 - A^p)$. The right-hand integral is dealt with in the same way, setting $A = 0$. Since the region of integration is $N$-dimensional, the measure is multiplied by $(1 - A^p)^N$. We solve for $A$ by solving $(1 - A^p)^N = 1/2$. We obtain $A = \left(1 - 2^{-1/N}\right)^{1/p}$, as required.

# 3 Ex.2.4

The main point is that $\sum \|x_i\|^2$ is invariant under the orthogonal group. As a consequence the standard normal distribution exists on any finite dimensional inner product space (by fixing an orthonormal basis). Further, if $R^p$ is written as the orthogonal sum of two vector subspaces, then the product of standard normal distributions on each of the subspaces gives the standard normal distribution on $R^p$. Everything else follows from this.

The on-line edition is correct except that $\sqrt{10} \approx 3.162278$. So, the figure given should be 3.2 instead of 3.1. Note that the version of this question posed in the first edition makes incorrect claims. The first edition talks of the "center of the training points" and the on-line edition talks of the "origin". The answers are very different. This is shown up best by taking only one training point.

# 4 Ex.2.7

## 4.1 (a)

For linear regression, we have

$$\hat{f}(x_0) = [x_0, 1] \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T y,$$

so that

$$\ell_i(x_0; \mathcal{X}) = [x_0, 1] \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \begin{pmatrix} 1 \\ x_i \end{pmatrix}$$

For $k$-nearest-neighbor regression, we have

$$\hat{f}(x_0) = \frac{1}{k} \sum_{x_1 \in N_k(x_0)} y_i,$$

where $N_k(x_0)$ is the neighborhood of $x_0$ defined by the $k$ closest points $x_i$ in the training sample.

$$\ell_i(x_0; \mathcal{X}) = \begin{cases} \frac{1}{k}, & \text{if } x_i \in N_k(x_0) \\ 0, & \text{otherwise.} \end{cases}$$

## 4.2 (b)

Note that $\mathcal{X}$ is fixed and randomness comes from $\mathcal{Y}$ only. We have

$$\begin{aligned} E_{y|\chi}\left(f(x_0) - \hat{f}(x_0)\right)^2 =& f(x_0)^2 - 2f(x_0) E_{y|\chi}\left(\hat{f}(x_0)\right) + E_{y|\chi}\left(\hat{f}(x_0)^2\right) \\ =& \left(f(x_0) - E_{y|\chi}\left(\hat{f}(x_0)\right)\right)^2 \\ &+ E_{y|\chi}\left(\hat{f}(x_0)^2\right) - \left(E_{y|\chi}\left(\hat{f}(x_0)\right)\right)^2 \\ =& \operatorname{Bias} y \mid \chi\left(\hat{f}(x_0)\right)^2 + \operatorname{Var}_{y|\chi}\left(\hat{f}(x_0)\right). \end{aligned}$$

### 4.3 (c)

The calculation logic is the same as (b), we have

$$\begin{aligned}
Ey, x\left(f\left(x_0\right)-\hat{f}\left(x_0\right)\right)^2 = & f\left(x_0\right)^2-2 f\left(x_0\right) Ey, x\left(\hat{f}\left(x_0\right)\right)+Ey, x\left(\hat{f}\left(x_0\right)^2\right) \\
= & \left(f\left(x_0\right)-Ey, \chi\left(\hat{f}\left(x_0\right)\right)\right)^2 \\
& +Ey, \chi\left(\hat{f}\left(x_0\right)^2\right)-\left(Ey, \chi\left(\hat{f}\left(x_0\right)\right)\right)^2 \\
= & \operatorname{Bias}\left(\hat{f}\left(x_0\right)\right)^2+\operatorname{Var}\left(\hat{f}\left(x_0\right)\right) .
\end{aligned}$$

### 4.4 (d)

From (b) we already see that $\operatorname{Bias} y \mid X\left(\hat{f}\left(x_0\right)\right)$ can be written as

$$\begin{aligned}
& f\left(x_0\right)-E_{y \mid X} \hat{f}\left(x_0\right) \\
= & f\left(x_0\right)-\sum_{i=1}^N E_Y\left\{\ell_i\left(x_0 ; \mathcal{X}\right)\left(f\left(x_i\right)+\epsilon_i\right)\right. \\
= & f\left(x_0\right)-\sum_{i=1}^N \ell_i\left(x_0 ; \mathcal{X}\right) f\left(x_i\right)
\end{aligned}$$

Also, we write $\operatorname{Var} y \mid x\left(\hat{f}\left(x_0\right)\right)$ as

$$\begin{aligned}
& E_y \mid \mathcal{X}\left(\hat{f}\left(x_0\right)^2\right)-\left(E_y, \mathcal{X}\left(\hat{f}\left(x_0\right)\right)\right)^2 \\
= & E_y \mid \mathcal{X}\left[\sum_{i=1}^N \sum_{j=1}^N \ell_i\left(x_0 ; \mathcal{X}\right) \ell_j\left(x_0 ; \mathcal{X}\right)\left(f\left(x_0\right)+\epsilon_i\right)\left(f\left(x_0\right)+\epsilon_j\right)\right] \\
& -\left(\sum_{i=1}^N \ell_i\left(x_0 ; \mathcal{X}\right) f\left(x_i\right)\right)^2 \\
= & \sum_{i=1}^N \sum_{j=1}^N \ell_i\left(x_0 ; \mathcal{X}\right) \ell_j\left(x_0 ; \mathcal{X}\right) f\left(x_i\right) f\left(x_j\right) \\
& +\sigma^2 \sum_{i=1}^N \ell_i^2\left(x_0 ; \mathcal{X}\right) \\
& -\sum_{i=1}^N \sum_{j=1}^N \ell_i\left(x_0 ; \mathcal{X}\right) \ell_j\left(x_0 ; \mathcal{X}\right) f\left(x_i\right) f\left(x_j\right) \\
= & \sigma^2 \sum_{i=1}^N \ell_i^2\left(x_0 ; \mathcal{X}\right)
\end{aligned}$$

Denote $S=\left(\ell_1\left(x_0 ; \mathcal{X}\right), \ldots, \ell_N\left(x_0 ; \mathcal{X}\right)\right)^T$ and $f=\left(f\left(x_1\right), \ldots, f\left(x_N\right)\right)^T$. By (2) and the equatior we have

$$\begin{aligned}
& \operatorname{Bias} y \mid x\left(\hat{f}\left(x_0\right)\right)=f\left(x_0\right)-S^T f, \\
& \operatorname{Vary} x\left(\hat{f}\left(x_0\right)\right)=\sigma^2 S^T S .
\end{aligned}$$

Assume that $SS^T$ is non-singular, note that $S^TS$ is a scalar, we have

$$\begin{aligned}\Big[\text{ Biasy }_{|x}\left(\hat{f}\left(x_0\right)\right)\Big]^2 = & \left(f\left(x_0\right) - S\left(x_0\right)^T f\right)^T \left(f\left(x_0\right) - S\left(x_0\right)^T f\right) \\ = & f\left(x_0\right)^2 + 2f\left(x_0\right) S^T f - f^T S S^T f \\ = & f\left(x_0\right)^2 + 2f\left(x_0\right) S^T f - f^T S S^T S S^T \left(SS^T\right)^{-1} f \\ = & f\left(x_0\right)^2 + 2f\left(x_0\right) S^T f - \frac{\text{Var}_{y|\chi}\left(\hat{f}\left(x_0\right)\right)}{\sigma^2} f^T f \\ = & f\left(x_0\right)^2 + 2f\left(x_0\right)\left(f\left(x_0\right) - \text{Biasy}|\chi\left(\hat{f}\left(x_0\right)\right)\right) \\ & - \frac{f^T f}{\sigma^2}\text{Var}_{\mathcal{Y}|\chi}\left(\hat{f}\left(x_0\right)\right).\end{aligned}$$

That is the relationship between the squared biases and variances. For (c), similar arguments follow by integrating terms above by joint density of $x_1, \ldots, x_N$, that is, $h\left(x_1\right)\cdots h\left(x_N\right)dx_1\cdots dx_N$