1.
$$||t_k - \widehat{y}|| = \sqrt{\sum_{i \neq k} \widehat{y}_i^2 + (1 - 2\widehat{y}_k + \widehat{y}_k^2)} = \sqrt{\sum_i \widehat{y}_i^2 + (1 - 2\widehat{y}_k)}.$$

In the right hand side, the first term $\sum_i \widehat{y}_i^2$ is independent of k, so we need only consider the second term, which is minimized when $\widehat{y}_k$ is maximized. Thus, $\text{argmin}||t_k - \widehat{y}||$ is equal to the k such that $\widehat{y}_k$ is maximized.

2. Let the orange means have corresponding pdfs o_i and the blue means have corresponding pdfs b_i. Then for any individual mean m_*, the probability that a particular point $p = (x_0, y_0)$ was sampled from the corresponding distribution D_* is $P(D_*|p) = P(p|D_*)P(D_*)/(\sum_i P(p|D_i)P(D_i))$. Since the means are sampled with equal probabilities and there are an equal number of blue and orange points, all the prior probabilities $P(D_i)$ are equal so the above reduces to $P(D_*|p) = P(p|D_*)/(\sum_i P(p|D_i)) = f_*(p)/\sum_i f_i(p)$. Summing over all the orange means, the probability that a point p is orange is $\sum_i o_i(p)/\sum_i o_i(p) + b_i(p)$ and we classify p as orange if this probability is larger than $1/2$. This yields the Bayes decision boundary where this probability is equal to $1/2$ (or equivalently $\sum_i o_i(p) = \sum_i b_i(p)$).

3. For a uniformly distributed point in the p-dimensional unit sphere, the cdf of the distance to the origin is $P(R < r) = cr^p/c1^p = r^p$. For N uniformly (independently) distributed points in the p-dimensional unit sphere, the minimum distance thus has cdf $P(\min(R_i) < r) = 1 - P(\min(R_i) \geq r) = 1 - P(R_1 \geq r)^N = 1 - (1 - r^p)^N$. The median of this distribution occurs when $P(\min(R_i) < r) = 1/2$, so solving $1/2 = 1 - (1 - r^p)^N$ for r yields $r = (1 - (1/2)^{1/N})^{1/p}$.

4. Since a is a unit vector, $||a||^2 = 1$. Since $x_i \sim N(0, I_p)$, each component $(x_i)_j$ is an independent $N(0,1)$ random variable and thus by properties of the normal distribution, $a^T x_i \sim N(\sum_i a_i 0 = 0, \sum_i a_i^2 1) = N(0, ||a||^2) = N(0,1)$. Thus $E(z_i^2) = Var(z_i) = 1$.

5. (a) In the prediction situation where the true relationship $Y = X^T\beta + \varepsilon$ is linear (up to noise $\varepsilon \sim N(0, \sigma^2)$), there are two sources of error: first, y_0 is not deterministically determined by x_0 so there is error coming from the model itself and second, the estimation of β depends on the training data we see.

   Thus, $EPE(x_0) = E((y_0 - \widehat{y_0})^2)$ where the expectation is over both $y_0(given x_0)$ and the training data T. (Note that although X is typically considered fixed, we can also view the training data T as randomly sampled pairs (X, Y) from some underlying joint distribution. Furthermore, $y_0|x_0$ and T are independent since the training and testing data are independent, so the joint expectation can be written successively as $E_{y_0|x_0}E_T$.) Adding and subtracting $x_0^T\beta$ yields $E((y_0 - x_0^T\beta)^2 + 2(y_0 - x_0^T\beta)(x_0^T\beta - \widehat{y_0}) + (x_0^T\beta - \widehat{y_0})^2)$. This further simplifies to $E(\varepsilon_0) + 2E(\varepsilon_0(x_0^T\beta - \widehat{y_0})) + E((x_0^T\beta - \widehat{y_0})^2)$. The first term does not depend on T, so the expectation over T is trivial and $E(\varepsilon_0^2) = \sigma^2$ by definition of the error term. (More precisely, $E((y_0 - x_0^T\beta)^2) = E_{y_0|x_0}((y_0 - x_0^T\beta)^2) = Var_{y_0|x_0}(y_0 - x_0^T\beta) + E_{y_0|x_0}(y_0 - x_0^T\beta)^2 = Var_{y_0|x_0}(y_0) = \sigma^2$.) Next, since the error term $\varepsilon_0 = y_0 - x_0^T\beta$ does not depend on T while the expectation $x_0^T\beta - \widehat{y_0}$ does not depend on y_0, the second term factors as $E_{y_0|x_0}(\varepsilon_0)E_T(x_0^T\beta - \widehat{y_0})$. By the model definition, both terms in the product above are zero. Thus we are left with $\sigma^2 + E_T((x_0^T\beta - x_0^T\widehat{\beta})^2)$. Focusing on the remaining expectation, it is equal to $Var_T(x_0^T\beta - x_0^T\widehat{\beta}) + E_T(x_0^T\beta - x_0^T\widehat{\beta})^2 = Var_T(x_0^T\widehat{\beta}) + (x_0^T\beta - E_T(x_0^T\widehat{\beta}))^2$. Since $\widehat{\beta}$ is unbiased when the model is linear, the latter term is zero (since it is the squared bias of the estimator $x_0^T\widehat{\beta}$ for the estimand $x_0^T\beta$). Simplifying the variance further:

$$Var_T(x_0^T\widehat{\beta}) = x_0^T Var_T((X^TX)^{-1}X^TY)x_0 = x_0^T(X^TX)^{-1}X^T Var_T(Y)X(X^TX)^{-1}x_0$$
$$= x_0^T(X^TX)^{-1}X^T\sigma^2 IX(X^TX)^{-1}x_0 = x_0^T(X^TX)^{-1}x_0\sigma^2$$

under the assumption that the errors are uncorrelated and viewing X as fixed. If we do not view X as fixed, then

$$Var_T((X^TX)^{-1}X^TY)$$
$$= E_T((X^TX)^{-1}X^TYY^TX(X^TX)^{-1}) - E_T((X^TX)^{-1}X^TY)E_T(Y^TX(X^TX)^{-1})$$
$$= E_T((\beta + (X^TX)^{-1}X^T\varepsilon)(\beta^T + \varepsilon^TX(X^TX)^{-1}))$$
$$- E_T(\beta + (X^TX)^{-1}X^T\varepsilon)E_T(\beta^T + \varepsilon^TX(X^TX)^{-1})$$
$$= E_T(\beta\beta^T) + E_T((X^TX)^{-1}X^T\varepsilon\beta^T) + E_T(\beta\varepsilon^TX(X^TX)^{-1}) + E_T((X^TX)^{-1}X^T\varepsilon\varepsilon^TX(X^TX)^{-1})$$
$$- E_T(\beta)E_T(\beta^T) - E_T((X^TX)^{-1}X^T\varepsilon)E_T(\beta^T) - E_T(\beta)E_T(\varepsilon^TX(X^TX)^{-1})$$
$$- E_T((X^TX)^{-1}X^T\varepsilon)E_T(\varepsilon^TX(X^TX)^{-1}).$$

β is fixed (but unknown) and in particular is constant with respect to T so can be treated as a constant in all the expectations above. By Adam's law $E_T((X^TX)^{-1}X^T\varepsilon) = E_T((X^TX)^{-1}X^TE_T(\varepsilon|X)) = 0$ since $E(\varepsilon|X) = 0$. After cancellation, the above becomes

$$= E_T((X^TX)^{-1}X^T\varepsilon\varepsilon^TX(X^TX)^{-1}) = E_T((X^TX)^{-1}X^TE_T(\varepsilon\varepsilon^T|X)X(X^TX)^{-1})$$
$$= E_T((X^TX)^{-1}X^T\sigma^2IX(X^TX)^{-1}) = E_T(X^TX^{-1})\sigma^2.$$

(This result seems to be inconsistent with respect to assuming that X is fixed.)

(b) Continuing not to assume that X is fixed, we now further assume that $E(X) = 0$ by centering (the columns of) X if necessary. Under this assumption, X^TX converges to $NCov(X)$ by the law of large numbers. (Otherwise, it would converge to $NCov(X) - E(X^T)E(X)$.) Thus, $(X^TX)^{-1}$ converges to $Cov(X)^{-1}/N$ (note that we do need the law of large numbers here, since even though $E_T(X^TX) = NCov(X)$, expectation is only linear so in particular does not necessarily interact well with inverses: it will not be true in general that as a result, $E_T((X^TX)^{-1}) = Cov(X)^{-1}/N$). Applying this to the result of the previous part,

$$E_{x_0}(\sigma^2 + x_0^TE_T((X^TX)^{-1})x_0\sigma^2) = \sigma^2 + \sigma^2/NE_{x_0}(x_0^TCov(X)^{-1}x_0)$$
$$= \sigma^2 + \sigma^2/NTr(E_{x_0}(x_0x_0^TCov(X)^{-1})) = \sigma^2 + \sigma^2/NTr(Cov(X)^{-1}E_{x_0}(x_0x_0^T))$$
$$= \sigma^2 + \sigma^2/NTr(Cov(X)^{-1}Cov(X)) = \sigma^2 + \sigma^2p/N.$$

6. Let x_i be only the distinct values of the inputs x_i and y_{ij} be all the output values corresponding to the input x_i. Then the (unweighted) RSS is $\sum_i \sum_j (y_{ij} - f_\theta(x_i))^2$. Let m_i be the number of times that input x_i is repeated. Then expanding the inner sum yields

$$\sum_j (y_{ij} - f_\theta(x_i))^2 = \sum_j y_{ij}^2 - 2\sum_j y_{ij} f_\theta(x_i) + m_i f_\theta(x_i)^2 = \sum_j y_{ij}^2 - 2m_i \overline{y_i} f_\theta(x_i) + m_i f_\theta(x_i)^2$$

$$= (\sum_j y_{ij}^2 - m_i \overline{y_i}^2) + m_i(\overline{y_i}^2 - 2\overline{y_i} f_\theta(x_i) + f_\theta(x_i)) = (\sum_j y_{ij}^2 - m_i \overline{y_i}^2) + m_i(\overline{y_i} - f_\theta(x_i))^2$$

and returning to the overall sum

$$\sum_i \sum_j (y_i j - f_\theta(x_i))^2 = \sum_i (\sum_j y_{ij}^2 - m_i \overline{y_i}^2) + m_i(\overline{y_i} - f_\theta(x_i))^2$$

$$= \sum_i (\sum_j y_{ij}^2 - m_i \overline{y_i}^2) + \sum_i m_i(\overline{y_i} - f_\theta(x_i))^2.$$

The second term on the RHS above is a weighted least squares problem and the first term does not depend on θ. In particular, minimizing the RSS $\sum_i \sum_j (y_{ij} - f_\theta(x_i))^2$ with respect to θ is equivalent to minimizing $\sum_i (\sum_j y_{ij}^2 - m_i \overline{y_i}^2) + \sum_i m_i(\overline{y_i} - f_\theta(x_i))^2$ with respect to θ. Since the first term does not depend on θ, this is further equivalent to minimizing the weighted least squares RSS $\sum_i m_i(\overline{y_i} - f_\theta(x_i))^2$ with respect to θ.

7. (a) For linear regression, the estimator $\widehat{f}(x_0)$ has least squares estimate $x_0^T(X^TX)^{-1}X^TY$. $x_0^T(X^TX)^{-1}X^T$ is a $1 \times N$ row vector whose entries depend only on x_0 and $\mathcal{X}$, so the inner product $(x_0^T(X^TX)^{-1}X^T)Y$ satisfies the required format.

   For kNN, ℓ_i is $1/k$ if x_i is one of the closest k neighbors to x_0 and 0 otherwise. Again, this function depends only on x_0 and $\mathcal{X}$ so it satisfies the required format.

   (b) $E_{Y|X}((f(x_0) - \widehat{f}(x_0))^2) = Var_{Y|X}(\widehat{f}(x_0)) + E_{Y|X}(f(x_0) - \widehat{f}(x_0))^2 = Var_{Y|X}(\widehat{f}(x_0)) + (f(x_0) - E_{Y|X}(\widehat{f}(x_0)))^2$ since $f(x_0)$ does not depend on the training data so can be considered constant with respect to the expectations/variances.

   (c) Using exactly analogous reasoning as the previous part, we obtain $Var_{Y,X}(\widehat{f}(x_0)) + (f(x_0) - E_{Y,X}(\widehat{f}(x_0)))^2$.

   (d) By Adam's law, $E_{Y,X}(-) = E_X(E_{Y|X}(-))$ so the (c) result is the expectation over X of the (b) result.

8.

9. Let $\tilde{\beta}$ be the least squares estimate for the test data. Since the testing and training data are drawn from the same underlying distribution/population, the least squares estimates $\widehat{\beta}$ and $\tilde{\beta}$ are identically distributed. (In particular, extra data points do not result in scaling of either estimate.) Thus all the terms $S_i = (y_i - \beta^T x_i)^2$ and $\tilde{S}_i = (\tilde{y}_i - \beta^T \tilde{x}_i)^2$ in the sums in R_{tr} and R_{te} are identically distributed (though not

independent) with expectation $E(S_i) = E(\tilde{S}_i) = S$ for some S. Thus $E(R_{tr}(\widehat{\beta})) = E(1/N \sum_i S_i) = 1/N \sum_i E(S_i) = S = 1/M \sum_i E(\tilde{S}_i) = E(1/M \sum_i \tilde{S}_i) = E(R_{te}(\tilde{\beta}))$.

Next, by definition of least squares, $R_{te}(\tilde{\beta})$ is minimized at $\tilde{\beta}$ and in particular $R_{te}(\tilde{\beta}) \leq R_{te}(\widehat{\beta})$. By monotonicity of expectation, $E(R_{te}(\tilde{\beta})) \leq E(R_{te}(\widehat{\beta}))$. Thus $E(R_{tr}(\widehat{\beta})) = E(R_{te}(\tilde{\beta})) \leq E(R_{te}(\widehat{\beta}))$