For myself:

$$\begin{aligned}
\mathrm{E}_{Y|X}\left(Y-c\right)^2 &= \mathrm{E}_{Y|X}\left(Y^2-2cY+c^2\right) && (1)\\
&= \mathrm{E}_{Y|X}Y^2-2c\,\mathrm{E}_{Y|X}Y+c^2 && (2)\\
&= \left[\mathrm{E}_{Y|X}Y^2-\left(\mathrm{E}_{Y|X}Y\right)^2\right]+\left(\mathrm{E}_{Y|X}Y-c\right)^2, && (3)
\end{aligned}$$

whence it follows that

$$\begin{aligned}
\arg\min \mathrm{E}_{Y|X}\left(Y-c\right)^2 &= \mathrm{E}_{Y|X}Y, && (4)\\
\min \mathrm{E}_{Y|X}\left(Y-c\right)^2 &= \mathrm{E}_{Y|X}Y^2-\left(\mathrm{E}_{Y|X}Y\right)^2 && (5)\\
&= \mathrm{Var}(Y|X). && (6)
\end{aligned}$$

Ex. 2.1: Suppose each of K-classes has an associated target t_k, which is a vector of all zeros, except a one in the kth position. Show that classifying to the largest element of $\hat{y}$ amounts to choosing the closest target, $\min_k \|t_k-\hat{y}\|$, if the elements of $\hat{y}$ sum to one.

Answer:

Let $\hat{y}\in\mathbb{R}^p : \sum_i \hat{y}_i = 1$. Then, for any j,

$$\begin{aligned}
\|t_j-\hat{y}\|^2 &= \sum_{i\neq j}\hat{y}_i^2+\left(1-\hat{y}_j\right)^2 && (7)\\
&= \sum_{i\neq j}\hat{y}_i^2+\left(\sum_{i\neq j}\hat{y}_i\right)^2 && (8)\\
&= 2\sum_{h,i\neq j;h\leq i}\hat{y}_h\hat{y}_i && (9)\\
&= 2\sum_{h\leq i}\hat{y}_h\hat{y}_i-2\sum_{h=j\vee i=j,h\leq i}\hat{y}_h\hat{y}_i && (10)\\
&= 2\sum_{h\leq i}\hat{y}_h\hat{y}_i-2\sum_{i}\hat{y}_k\hat{y}_i && (11)\\
&= 2\sum_{h\leq i}\hat{y}_h\hat{y}_i-2\hat{y}_k\sum_{i}\hat{y}_i && (12)\\
&= 2\sum_{h\leq i}\hat{y}_h\hat{y}_i-2\hat{y}_k && (13)\\
& && (14)
\end{aligned}$$

Soo, if the closest thingy is the kth,

$$\begin{aligned}
\|t_k-\hat{y}\| &\leq \|t_j-\hat{y}\|\forall j, && (15)\\
\|t_k-\hat{y}\|^2 &\leq \|t_j-\hat{y}\|^2\forall j, && (16)\\
2\sum_{h\leq i}\hat{y}_h\hat{y}_i-2\hat{y}_k &\leq 2\sum_{h\leq i}\hat{y}_h\hat{y}_i-2\hat{y}_j\forall j, && (17)\\
\hat{y}_k &\geq \hat{y}_j\forall j. && (18)
\end{aligned}$$

Ex. 2.2: Show how to compute the Bayes decision boundary for the simulation example in Figure 2.5.

Answer:

Maybe the previous excercise was meant to help with this one, since we have to pick k such that $p_k \geq p_j\forall j$ with p in the probability simplex, which is equivalent to picking the axis closest to p. But that doesn't look less

computationaly expensive. In the general case, I guess you could construct a grid and simply check for those conditions.

In the case of a binary classification problem, you can take advantage in the fact that the decision boundary takes the form $p_1 = 0.5$, and just move along the unique choice perpendicular to the gradient of p_1.

Ex. 2.3: Derive equation (2.24).

Answer:

We have a uniform probability distribution along $S_p = \{x \in \mathbb{R}^p : \|x\| \leq 1\}$, and we take N samples. The median minimum distance is such that the probability of a smaller minimum distance be $1/2$. The probability that the minimum distance is above r is the probability that all N samples are above r, which should be the Nth power of the probability that any individual sample is above r, which is $1 - r^p$. That is,

$$P(\|x\| \geq r) = (1 - r^p)^N \tag{19}$$

and, setting this to equal $1/2$, we get

$$d(p,\, N) = \left(1 - \frac{1}{2^{1/N}}\right)^{1/p}. \tag{20}$$

Thank God for commulative probabilities.

Ex. 2.4: The edge effect problem discussed on page 15 is not peculiar to uniform sampling from bounded domains. Consider inputs drawn from a spherical multinormal distribution $X \sim N(0,\, \mathbf{I}_p)$. The squared distance from any sample point to the origin has a χ^2_p distribution with mean p. Consider a prediction point x_0 drawn from this distribution, and let $a = x_0/\|x_0\|$ be an associated unit vector. Let $z_i = a^{\mathrm{T}} x_i$ be the projection of each of the training points in this direction.

a Show that the z_i are distributed $N(0,\, 1)$ with expected squared distance from the origin 1, while the target point has expected squared distance p from the origin.

b For $p = 10$ show that the expected distance of a test point from the center of the training data is 3.1 standard deviations, while all the training points have expected distance 1 along direction a. So most prediction points see themselves as lying on the edge of the training set.

Answer:

Part a:

$N(0,\, \mathbf{I}_p)$ is precisely obtained as the probability distribution of a point each of whose components are taken from $N(0,\, 1)$, which shows $a^T x_i$ is taken from a distribution $N(0,\, 1)$.

Long road: Let $P_X = N(\mu,\, \Sigma)$, which is to say

$$P_X(x) = \frac{1}{\sqrt{2\pi \det \Sigma^p}} \exp\left(-\frac{1}{2}\left(x - \mu\right)^T \Sigma^{-1} \left(x - \mu\right)\right). \tag{21}$$

Then,

$$P_{a^T X}(y) = \int_{\mathbb{R}^p} \mathrm{d}x\, P_X(x) \delta(a^T x - y) \tag{22}$$

$$= \int_{\mathbb{R}^p :\, a^T x = y} \mathrm{d}x\, P_X(x) \frac{1}{\cancelto{1}{\|a\|}} \tag{23}$$

We now take into account that $a^T x = y$ means $x = \sum_i x_i u_i + ya = x_\perp + ya$ for $\{a, u_i\}$ an orthonormal basis. This way, and decomposing $\mu = \mu_\perp + \mu_a a$,

$$(x - \mu)^T \Sigma^{-1} (x - \mu) = (x_\perp - \mu_\perp + (y - \mu_a)a)^T \Sigma^{-1} (x_\perp - \mu_\perp + (y - \mu_a)a) \tag{24}$$

$$= (x_\perp - \mu_\perp)^T \Sigma^{-1} (x_\perp - \mu_\perp) \tag{25}$$

$$+ 2(x_\perp - \mu_\perp)^T \Sigma^{-1} (y - \mu_a)a \tag{26}$$

$$+ (y - \mu_a)^2 a^T \Sigma^{-1} a. \tag{27}$$

Aaand so we plug all that into the big integral. The (exponential of the) last term can be taken outside the integral, whereas the other two represent stuff to integrate. The contributions from $x_\perp - \mu_\perp$ and $\mu_\perp - x_\perp$ cancel out the second factor, and so it ends up being a constant factor. We can then conclude the probability distribution for y is $N(\mu_a,\, a^T\Sigma^{-1}a)$.

On the other hand, the expected square distance $\mathrm{E}\|x\|^2$ is given by

$$\mathrm{E}\|x\|^2 = \frac{\int \mathrm{d}x\, \|x\|^2 \exp\left(-\frac{1}{2}\|x\|^2\right)}{\int \mathrm{d}x\, \exp\left(-\frac{1}{2}\|x\|^2\right)} \tag{28}$$

$$= \frac{\int_0^\infty \mathrm{d}r\, r^{p-1} \int \mathrm{d}\Theta\, r^2 \exp\left(-\frac{1}{2}r^2\right)}{\int_0^\infty \mathrm{d}r\, r^{p-1} \int \mathrm{d}\Theta\, \exp\left(-\frac{1}{2}r^2\right)} \tag{29}$$

$$= \frac{\int_0^\infty \mathrm{d}r\, r^{p+1} \exp\left(-\frac{1}{2}r^2\right)}{\int_0^\infty \mathrm{d}r\, r^{p-1} \exp\left(-\frac{1}{2}r^2\right)} \tag{30}$$

$$= \frac{\int_0^\infty \mathrm{d}\left(\frac{1}{2}r^2\right)\, 2^{\frac{1}{2}p} \left(\frac{1}{2}r^2\right)^{\frac{1}{2}p} \exp\left(-\frac{1}{2}r^2\right)}{\int_0^\infty \mathrm{d}\left(\frac{1}{2}r^2\right)\, 2^{\frac{1}{2}p-1} \left(\frac{1}{2}r^2\right)^{\frac{1}{2}p-1} \exp\left(-\frac{1}{2}r^2\right)} \tag{31}$$

$$= 2\frac{\int_0^\infty \mathrm{d}t\, t^{\frac{1}{2}p} \exp(-t)}{\int_0^\infty \mathrm{d}t\, t^{\frac{1}{2}p-1} \exp(-t)} \tag{32}$$

$$= 2\frac{\Gamma\left(\frac{1}{2}p+1\right)}{\Gamma\left(\frac{1}{2}p\right)} \tag{33}$$

$$= p. \tag{34}$$

Note to self: Gamma thingy can be done through parts:

$$\Gamma(x+1) = \int_0^\infty \mathrm{d}t\, t^x \exp(-t) \tag{35}$$

$$= \underbrace{\left[-xt^{x-1}\exp(-t)\right]_0^\infty}_{\to 0} + \int_0^\infty \mathrm{d}t\, xt^{x-1} \exp(-t) \tag{36}$$

$$= x\Gamma(x) \tag{37}$$

Part b:

If we compute the expected distance as $\sqrt{\mathrm{E}\|x\|^2}$ we're left with about 3.16, which I guess is what the author did. And here I was, about to compute $\mathrm{E}\|x\|$. In fact, let's do it for the lulz.

$$\mathrm{E}\|x\| = \frac{\int \mathrm{d}r\, \|x\| \exp\left(-\frac{1}{2}\|x\|^2\right)}{\int \mathrm{d}x\, \exp\left(-\frac{1}{2}\|x\|^2\right)} \tag{38}$$

$$= \cdots \tag{39}$$

$$= \frac{\int \mathrm{d}r\, r^p \exp\left(-\frac{1}{2}r^2\right)}{\int \mathrm{d}x\, r^{p-1} \exp\left(-\frac{1}{2}r^2\right)} \tag{40}$$

$$= \sqrt{2}\frac{\Gamma\left(\frac{1}{2}p+\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}p\right)} \tag{41}$$

$$\approx 3.08. \tag{42}$$

For the expected distance along direction a, we get 1 as the square root of the expected square distance, and

$$\mathrm{E}\|x\| = \frac{\int_0^\infty \mathrm{d}r\, r \exp\left(-\frac{1}{2}r^2\right)}{\int_0^\infty \mathrm{d}r\, \exp\left(-\frac{1}{2}r^2\right)} \tag{43}$$

$$= \frac{\int_0^\infty \mathrm{d}t\, \exp(-t)}{\sqrt{\frac{\pi}{2}}} \tag{44}$$

$$= \frac{2}{\pi} \tag{45}$$

$$\approx 0.78. \tag{46}$$

Ex. 2.5:

a Derive equation (2.27). The last line makes use of (3.8) through a conditioning argument.

b Derive equation (2.28), making use of the *cyclic* property of the trace operator [$\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$], and its linearity (which allows to interchange the order of trace and expectation).

Answer:

The relationship between the random variables X and Y is

$$Y = X^T\beta + \varepsilon, \tag{47}$$

with $\varepsilon \sim N(0,\, \sigma)$ and X taking values in $\mathbb{R}^p$. As a training set, we have N inputs arranged in row form into matrix $\mathbf{X} \in \mathbb{R}^{N\times p}$, and N outputs aranged in $\mathbf{y} \in \mathbb{R}^{N\times 1}$ of the form $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is a vector of samples from ε's probability distribution function.

So, the expected prediction error at input value x_0 is given by

$$\mathrm{EPE}(x_0) = \mathrm{E}_{y_0|x_0}\mathrm{E}_{\mathcal{T}}\left(y_0 - \hat{y_0}\right)^2, \tag{48}$$

$$\mathrm{E}_{y_0|x_0} = \int \mathrm{d}y_0\, P(y_0|x_0), \tag{49}$$

$$\mathrm{E}_{\mathcal{T}} = \mathrm{E}_{\mathbf{X}}\mathrm{E}_{\varepsilon} \tag{50}$$

$$= \int \mathrm{d}\mathbf{X}\, P(\mathbf{X}) \int \mathrm{d}\boldsymbol{\varepsilon}\, P(\boldsymbol{\varepsilon}). \tag{51}$$

We observe that all y_0 (ε), $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ are independent variables, whence the expectation values commute.

The estimated value is defined by

$$\hat{y_0} = x_0^T\hat{\beta}, \tag{52}$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{53}$$

$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\beta + \boldsymbol{\varepsilon}) \tag{54}$$

$$= \beta + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}. \tag{55}$$

We thus have:

$$\begin{aligned}
\mathrm{E}_{y_0|x_0} y_0 &= \mathrm{E}_{\varepsilon|x_0}\left(x_0^T\beta + \varepsilon\right) && (56)\\
&= x_0^T\beta + \cancelto{0}{\mathrm{E}_{\varepsilon|x_0}\varepsilon}, && (57)\\
\mathrm{Var}(y_0|x_0) &= \mathrm{E}_{\varepsilon|x_0}\left(y_0 - \mathrm{E}_{\varepsilon|x_0} y_0\right)^2 && (58)\\
&= \mathrm{E}_{\varepsilon|x_0}\left(x_0^T\beta + \varepsilon - x_0^T\beta\right)^2 && (59)\\
&= \mathrm{E}_{\varepsilon|x_0}\varepsilon^2 = 1, && (60)\\
\mathrm{E}_{\mathcal{T}}\hat{y_0} &= \mathrm{E}_{\mathcal{T}}\left(x_0^T\hat{\beta}\right) && (61)\\
&= \mathrm{E}_{\mathcal{T}}\left(x_0^T\beta + \varepsilon\right) && (62)\\
&= x_0^T\beta + \cancelto{0}{\mathrm{E}_{\mathcal{T}}\varepsilon}, && (63)\\
\mathrm{Var}\hat{y_0} &= \mathrm{E}_{\mathcal{T}}\left(\hat{y_0} - \mathrm{E}_{\mathcal{T}}\hat{y_0}\right)^2 && (64)\\
&= \mathrm{E}_{\mathcal{T}}\left(x_0^T\beta + x_0^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\varepsilon} - x_0^T\beta\right)^2 && (65)\\
&= \mathrm{E}_{\mathcal{T}}\left(x_0^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\right)^2 && (66)\\
&= \mathrm{E}_{\mathcal{T}} x_0^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} x_0 && (67)\\
&= \mathrm{E}_{\mathbf{X}} x_0^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathrm{E}_{\boldsymbol{\varepsilon}}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right)\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} x_0 && (68)\\
&= \mathrm{E}_{\mathbf{X}} x_0^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\sigma^2\mathrm{I}_N\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} x_0 && (69)\\
&= \sigma^2\mathrm{E}_{\mathbf{X}} x_0^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1} x_0 && (70)\\
& && (71)
\end{aligned}$$

Now,

$$\begin{aligned}
\mathrm{EPE}(x_0) &= \mathrm{E}_{y_0|x_0}\mathrm{E}_{\mathcal{T}}\left(y_0 - \hat{y_0}\right)^2 && (72)\\
&= \mathrm{E}_{y_0|x_0} y_0^2 - 2\mathrm{E}_{y_0|x_0} y_0\mathrm{E}_{\mathcal{T}}\hat{y_0} + \mathrm{E}_{\mathcal{T}}\hat{y_0}^2 && (73)\\
&= \mathrm{Var}(y_0|x_0) + \left(\mathrm{E}_{y_0|x_0} y_0\right)^2 - 2\mathrm{E}_{y_0|x_0} y_0\mathrm{E}_{\mathcal{T}}\hat{y_0} + \mathrm{Var}_{\mathcal{T}}(\hat{y_0}) + \left(\mathrm{E}_{\mathcal{T}}\hat{y_0}\right)^2 && (74)\\
&= \mathrm{Var}(y_0|x_0) + \mathrm{Var}_{\mathcal{T}}(\hat{y_0}) + \left(\mathrm{E}_{y_0|x_0} y_0 - \mathrm{E}_{\mathcal{T}}\hat{y_0}\right)^2 && (75)\\
&= \mathrm{Var}(y_0|x_0) + \mathrm{Var}_{\mathcal{T}}(\hat{y_0}) + \mathrm{Bias}^2(\hat{y_0}) && (76)\\
&= \sigma^2 + \sigma^2 x_0^T\mathrm{E}_{\mathbf{X}}(\mathbf{X}^T\mathbf{X})^{-1} x_0 + 0. && (77)
\end{aligned}$$

Since

$$\mathbf{X}^T\mathbf{X} = \left(\sum_{\mu} x_i^{\mu} x_j^{\mu}\right), \tag{78}$$

and assuming $\mathrm{E}X = 0$ for each element, this is approximately $N\mathrm{Cov}(X)$. So,

$$\mathrm{E}_{x_0}\mathrm{EPE}(x_0) = \sigma^2 + \frac{\sigma^2}{N}\mathrm{E}_{x_0}\left(x_0^T\mathrm{Cov}^{-1}(X)x_0\right) \tag{79}$$

$$= \sigma^2 + \frac{\sigma^2}{N}\mathrm{E}_{x_0}\,\mathrm{Tr}\left(x_0^T\mathrm{Cov}^{-1}(X)x_0\right) \tag{80}$$

$$= \sigma^2 + \frac{\sigma^2}{N}\,\mathrm{Tr}\left(\mathrm{Cov}^{-1}(X)\mathrm{E}_{x_0}x_0x_0^T\right) \tag{81}$$

$$= \sigma^2 + \frac{\sigma^2}{N}\,\mathrm{Tr}\left(\mathrm{Cov}^{-1}(X)\mathrm{Cov}(x_0)\right) \tag{82}$$

$$= \sigma^2 + \frac{\sigma^2}{N}\,\mathrm{Tr}\left(\mathrm{I}_p\right) \tag{83}$$

$$= \sigma^2 + \sigma^2\frac{p}{N} \tag{84}$$

$$\tag{85}$$

Ex. 2.6: Consider a regression problem with imputs x_i and outputs y_i, and a parametrized model $f_\theta(x)$ to be fit by least squares. Show that if there are observations with *tied* or *identical* values of x, then the fit can be obtained from a reduced weighted least squares problem.

Answer:

Let $I_{i,j}$ be the index for the ith unique value X_i (from 1 to U), with j ranging from 1 to the number of repetitions R_i. Then,

$$\mathrm{RSS}(\theta) = \sum_{i=1}^{N}\left(y_i - f_\theta(x_i)\right)^2 \tag{86}$$

$$= \sum_{i=1}^{U}\sum_{j=1}^{R_i}\left(y_{I_{i,j}} - f_\theta(x_{I_{i,j}})\right)^2 \tag{87}$$

$$= \sum_{i=1}^{U}\sum_{j=1}^{R_i}\left(y_{I_{i,j}} - f_\theta(X_i)\right)^2 \tag{88}$$

$$= \sum_{i=1}^{U}\sum_{j=1}^{R_i}\left(y_{I_{i,j}}^2 - 2y_{I_{i,j}}f_\theta(X_i) + f_\theta^2(X_i)\right) \tag{89}$$

$$= \sum_{i=1}^{U}\sum_{j=1}^{R_i}\left(y_{I_{i,j}}^2 - 2y_{I_{i,j}}f_\theta(X_i) + f_\theta^2(X_i)\right) \tag{90}$$

$$= \sum_{i=1}^{U}R_i\left(-2\bar{y}_i f_\theta(X_i) + f_\theta^2(X_i)\right) + \sum_{i=1}^{U}\sum_{j=1}^{R_i}y_{I_{i,j}}^2 \tag{91}$$

$$= \sum_{i=1}^{U}R_i\left(\bar{y}_i - f_\theta(X_i)\right)^2 - \sum_{i=1}^{U}R_i\bar{y}_i^2 + \sum_{i=1}^{U}\sum_{j=1}^{R_i}y_{I_{i,j}}^2, \tag{92}$$

$$\bar{y}_i = \frac{1}{R_i}\sum_{j=1}^{R_i}y_{ij}. \tag{93}$$

Since the later two sumands are constant with respect to θ, the minimum of this function equals the minimum of

$$\mathrm{RSS}(\theta) = \sum_{i=1}^{U}R_i\left(\bar{y}_i - f_\theta(X_i)\right)^2. \tag{94}$$

Ex. 2.7: Suppose we have a sample of N pairs x_i, y_i drawn i.i.d. from the distribution characterized as follows:

- $x_i \sim h(x)$, the design density
- $y_i = f(x_i) + \varepsilon_i$, f is the regression function
- $\varepsilon_i \sim (0, \sigma^2)$ (mean zero, variance σ^2)

We construct an estimator for f *linear* in the y_i,

$$\hat{f}(x_0) = \sum_{i=1}^{N} l_i(x_0; \mathcal{X}) y_i \tag{95}$$

where the weights $l_i(x_0; \mathcal{X})$ do not depend on the y_i, but do depend on the entire training sequence of x_i, denoted here by $\mathcal{X}$.

a Show that linear regression and k-nearest-neighbor regression are members of this class of estimators. Describe explicitly the weights $l_i(x_0; \mathcal{X})$ for both these cases.

b Decompose the conditional mean-squared error

$$\mathrm{E}_{\mathcal{Y}|\mathcal{X}} \left(f(x_0) - \hat{f}(x_0) \right)^2 \tag{96}$$

into a conditional squared bias and a conditional variance component. Like $\mathcal{X}$, $\mathcal{Y}$ represents the entire training sequence of y_i.

c Decompose the (unconditional) mean-squared error

$$\mathrm{E}_{\mathcal{X},\mathcal{Y}} \left(f(x_0) - \hat{f}(x_0) \right)^2 \tag{97}$$

into a squared bias and a variance component.

d Establish a relationship between the squared biases and variances in the above two cases.

Answer:

Part a:

Linear regression prescribes

$$\hat{f}(x_0) = \mathbf{l}^T(x_0; \mathcal{X}) \mathbf{y} \tag{98}$$

$$= \sum_{i}^{N} l_i(x_0; \mathcal{X}) y_i, \tag{99}$$

$$\mathbf{l}(x_0; \mathcal{X}) = x_0^T \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T. \tag{100}$$

KNN prescribes

$$\hat{f}(x_0) = \sum_{i=1}^{N} l_i(x_0; \mathcal{X}) y_i, \tag{101}$$

$$l_i(x_0; \mathcal{X}) = \frac{1}{K} \chi_{\mathrm{KNN}(x_0)}(x_i), \tag{102}$$

$$\mathrm{KNN}(x_0): \text{ set of K nearest neighbors to } x_0. \tag{103}$$

Part b and c:

$$\mathrm{E} \left(f(x_0) - \hat{f}(x_0) \right)^2 = f^2(x_0) - 2f(x_0)\mathrm{E}\hat{f}(x_0) + \mathrm{E}\hat{f}^2(x_0) \tag{104}$$

$$= \left(f(x_0) - \mathrm{E}\hat{f}(x_0) \right)^2 + \mathrm{E}\hat{f}^2(x_0) - \left(\mathrm{E}f(x_0) \right)^2 \tag{105}$$

$$= \mathrm{Bias}^2 \left(\hat{f}(x_0) \right) + \mathrm{Var} \left(\hat{f}(x_0) \right), \tag{106}$$

since in both cases x_0 and $f(x_0)$ are constant.

Part d:

In part b, the sample set is constant, whereas in part c it is averaged. But we can decompose $\mathrm{E}_{\mathcal{X},\mathcal{Y}}$ as $\mathrm{E}_{\mathcal{X}}\mathrm{E}_{\mathcal{Y}|\mathcal{X}}$. Soo, for the bias, we have

$$\mathrm{Bias}_{\mathcal{X},\mathcal{Y}}\left(\hat{f}(x_0)\right) = \mathrm{E}_{\mathcal{X},\mathcal{Y}}\hat{f}(x_0) - f(x_0) \tag{107}$$

$$= \mathrm{E}_{\mathcal{X}}\mathrm{E}_{\mathcal{Y}|\mathcal{X}}\left(\hat{f}(x_0) - f(x_0)\right) \tag{108}$$

$$= \mathrm{E}_{\mathcal{X}}\mathrm{Bias}_{\mathcal{Y}|\mathcal{X}}\left(\hat{f}(x_0)\right). \tag{109}$$

Now, in this particular case,

$$\mathrm{Bias}_{\mathcal{Y}|\mathcal{X}}\left(\hat{f}(x_0)\right) = \mathrm{E}_{\mathcal{Y}|\mathcal{X}}\left(\hat{f}(x_0) - f(x_0)\right) \tag{110}$$

$$= \mathrm{E}_{\mathcal{Y}|\mathcal{X}}\left(\sum_i l_i(x_0|\mathcal{X})y_i - f(x_0)\right) \tag{111}$$

$$= \mathrm{E}_{\mathcal{Y}|\mathcal{X}}\left(\sum_i l_i(x_0|\mathcal{X})(f(x_i)+\varepsilon_i) - f(x_0)\right) \tag{112}$$

$$= \sum_i l_i(x_0|\mathcal{X})f(x_i) + \mathrm{E}_{\boldsymbol{\varepsilon}}\sum_i l_i(x_0|\mathcal{X})\varepsilon_i - f(x_0) \tag{113}$$

$$= \sum_i l_i(x_0|\mathcal{X})f(x_i) - f(x_0). \tag{114}$$

So, the conditional bias cannot be made null without knowledge of $f(x_0)$, since it depends on the particular samples drawn. However, KNN is unbiased when averaged over training samples. To see this, we note the leftmost term is simply the sum over the K nearest neighbors. For the region where the K nearest neighbors are the ones corresponding to the samples with index I, the indexes may be relabeled so that $I = [1,\, K]_{\mathbb{N}}$. There are K taken from N such regions. Of these, for the region where the ith index corresponds to the Kth closest neighbor, again the indexes may be relabeled so that the first one is the closest one. There are N such regions. So we have

$$\mathrm{E}_{\mathcal{X}}\sum_{i=1}^{N} l_i(x_0|\mathcal{X})f(x_i) = \prod_{i=1}^{N}\int_{-\infty}^{\infty}\mathrm{d}x_i\, h(x_i - x_0)\sum_{j=1}^{N} l_j(x_0|\mathcal{X})f(x_j - x_0) \tag{115}$$

$$= N\binom{N}{K}\int_0^{\infty}\mathrm{d}x_1\ (h(x_1)+h(-x_1))\prod_{i=2}^{K}\int_{-x_1}^{x_1}\mathrm{d}x_i\, h(x_i)\frac{1}{K}\sum_j f(x_j+x_0)\left(\int_{-\infty}^{-x_1}\mathrm{d}x\, h(x) + \int_{x_1}^{\infty}\mathrm{d}x\, h(x)\right) \tag{116}$$

Aaand I'm at a loss here.

Ex. 2.8: Compare the classification performance of linear regression and k-nearest neighbor classification on the zipcode data. In particular, consider only the 2's and 3's, and $k = 1$, 3, 5, 7 and 15. Show both the training and test error for each choice. The zipcode data are available from the book website www−stat.stanford.edu/ElemStatLearn.

Answer:

The code "2.8.py" produces figure 1. It's noticeable that linear regression performs worse with the test data, but generalizes better, in this case, than all other instances of KNN.

Ex. 2.9: Consider a linear regression model with p parameters, fit by least squares to a set of training data $(x_1, y_1), \ldots, (x_N, y_N)$ drawn at random from a population. Let $\hat{\beta}$ be the least squares estimate. Suppose we have some test data $(\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_M, \tilde{x}_M)$ drawn at random from the same population as the training data. If $R_{tr}(\beta) = \frac{1}{N}\sum_i^N \left(y_i - \beta^T x_i\right)^2$ and $R_{te}(\beta) = \frac{1}{M}\sum_i^M \left(\tilde{y}_i - \beta^T \tilde{x}_i\right)^2$, prove that

$$\mathrm{E}[R_{tr}(\hat{\beta})] \leq \mathrm{E}[R_{ts}(\hat{\beta})], \tag{117}$$

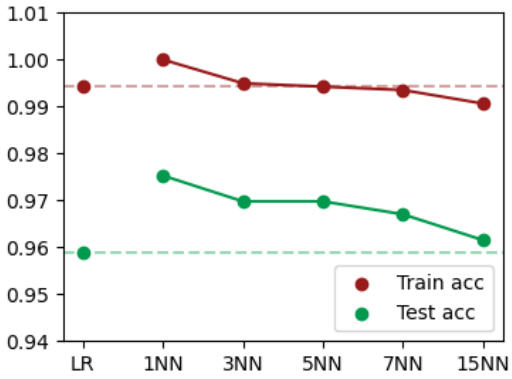

Figure 1: Accuracy as a function of the regressor. "LR" corresponds to Linear Regressor, and "iNN" corresponds to an i-nearest neighbors regressor.

where the expectations are over all that is random in each expression.

Answer:

Bear in mind that $\mathrm{E}f(x_i, y_i, \hat{\beta}) = \mathrm{E}f(x_j, y_j, \hat{\beta})$. To see this, first note $\hat{\beta}$ is invariant with respect to changes of labels:

$$\hat{\beta}(\mathbf{X}, \mathbf{y}) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{118}$$

$$= \left(\mathbf{X}^T S_{ij}^T S_{ij}\mathbf{X}\right)^{-1}\mathbf{X}^T S_{ij}^T S_{ij}\mathbf{y} \tag{119}$$

$$= \left((S_{ij}\mathbf{X})^T (S_{ij}\mathbf{X})\right)^{-1}(S_{ij}\mathbf{X})^T (S_{ij}\mathbf{y}) \tag{120}$$

$$= \hat{\beta}(S_{ij}\mathbf{X}, S_{ij}\mathbf{y}), \tag{121}$$

where $S_{ij} = S_{ij}^T = S_{ij}^{-1}$ is the i to j row operator. It then follows that

$$\mathrm{E}f(x_i, y_i, \hat{\beta}) = \int P(\mathrm{d}x_1, \mathrm{d}y_1)\dots P(\mathrm{d}x_N, \mathrm{d}y_N)\, f(x_i, y_i, \hat{\beta}) \tag{122}$$

$$= \int P(\mathrm{d}x_1, \mathrm{d}y_1)\dots P(\mathrm{d}x_N, \mathrm{d}y_N)\, f(x_j, y_j, \hat{\beta}) \text{ (Change of labels)} \tag{123}$$

$$= \mathrm{E}f(x_j, y_j, \hat{\beta}). \tag{124}$$

Thus,

$$\mathrm{E}[R_{tr}(\hat{\beta}(\mathbf{X}, \mathbf{y}))] = \mathrm{E}\left[\frac{1}{N}\sum_i^N \left(y_i - \hat{\beta}^T(\mathbf{X}, \mathbf{y})x_i\right)^2\right] \tag{125}$$

$$= \frac{1}{N}\sum_i^N \mathrm{E}\left(y_i - \hat{\beta}^T(\mathbf{X}, \mathbf{y})x_i\right)^2 \tag{126}$$

$$= \mathrm{E}\left(y_1 - \hat{\beta}^T(\mathbf{X}, \mathbf{y})x_1\right)^2, \tag{127}$$

$$\mathrm{E}[R_{ts}(\hat{\beta}(\mathbf{X}, \mathbf{y}))] = \mathrm{E}\left(\tilde{y}_1 - \hat{\beta}^T(\mathbf{X}, \mathbf{y})\tilde{x}_1\right)^2. \tag{128}$$

Now, for the comparison,

$$\mathrm{E}[R_{tr}(\hat{\beta}(\mathbf{X}, \mathbf{y}))] = \mathrm{E}\left(y_1 - \hat{\beta}^T(\mathbf{X}, \mathbf{y})x_1\right)^2 \tag{129}$$

$$= \mathrm{E}y_1^2 - 2\mathrm{E}y_1\hat{\beta}^T(\mathbf{X}, \mathbf{y})x_1 + \mathrm{E}\left[\hat{\beta}(\mathbf{X}, \mathbf{y})^T x_1\right]^2 \tag{130}$$

Again I'm at a loss. It's time to move on.