1. (a) $\frac{\partial}{\partial\beta}E[(Y-X^T\beta)^2]$

$= \frac{\partial}{\partial\beta}E[Y^TY - 2Y^TX^T\beta + (X^T\beta)^T \cdot X^T\beta] = E[-2XY + 2XX^T\beta] = 0$

$\therefore 2E[XX^T]\beta = 2E[XY]$

$\therefore E[XX^T]\beta = E[XY]$

$\therefore \beta = E^{-1}[XX^T]E[XY]$

(b) The nearest neighbors calculate the average number of the labels. of some points that are close to aim point x.
The least squares use the function $\beta = (X^TX)^{-1}X^Ty$ to get the parameters β to minimize the squared error loss and then approximate the regression function.
Difference: ① The nearest neighbors is local because for a new point x, we only consicler the points that is close to x. It has no consumption for model. When dimension grows, we need exponential increase of sample points to approximate the regression function. It's Bias will not change too much but Var will be higher when dimension grows.
② The least square is global because it considers all the sample points in the training set. It consumes the model linearly before training. When dimension grows, we need linear increase of sample points to approximate the regression function.

(c) $EPE(f) = E(|Y - f(x)|)$

$= E_X E_{Y|X}([Y - f(x)] \mid X)$

$\therefore f(x) = \underset{f}{\operatorname{argmin}} E_{Y|X}[|Y - f(x)| \mid X = x]$

$= \underset{f}{\operatorname{argmin}} \int_y |y - f(x)| \Pr(y|x)\, dy$

$= \underset{f}{\operatorname{argmin}} \left(\int_{y \geq f(x)} (y - f(x)) \Pr(y|x)\, dy + \int_{y < f(x)} (f(x) - y) \Pr(y|x)\, dy \right)$

$= \underset{f}{\operatorname{argmin}} \underbrace{\left(\int_{y \geq f(x)} y \Pr(y|x)\, dy - \int_{y < f(x)} y \Pr(y|x)\, dy + f(x) \Pr(y < f(x) \mid x) - f(x) \Pr(y \geq f(x) \mid x) \right)}_{A}$

$\therefore \frac{\partial A}{\partial f(x)} = \Pr(y < f(x) \mid x) - \Pr(y \geq f(x) \mid x)$

$= 2 \Pr(y < f(x) \mid x) - 1 = 0$

$\therefore \Pr(y < f(x) \mid x) = \frac{1}{2}$

$\therefore f(x) = \text{median}(Y \mid X = x)$

2.(a) $\|\hat{y}-\hat{X}w\|_2^2 = \sum_{i=1}^{n}(y_i - x_i w_i)^2 + \sum_{i=n+1}^{d+n}(y_i - x_i w_i)^2$

$= \sum_{n=1}^{n}(y_i - x_i w_i)^2 + \sum_{i=n+1}^{n+d}(\sqrt{\lambda} w_i)^2$

$= \|y - Xw\|_2^2 + \lambda\|w\|_2^2$

$\therefore \frac{\partial\|\hat{y}-\hat{X}w\|_2^2}{\partial w} = \frac{\partial(\|y-Xw\|_2^2 + \lambda\|w\|_2^2)}{\partial w} = \frac{\partial(y^Ty - 2y^TXw + w^TX^TXw + \lambda w^Tw)}{\partial w}$

$= -2X^Ty + 2X^TXw + 2\lambda w = 0$

$\therefore (X^TX + \lambda I_d)w = X^Ty$

$\therefore w = (X^TX + \lambda I_d)^{-1}X^Ty$

(b) $\hat{X}\beta = y \quad \therefore \beta \in R^{n+d}$

Suppose $\beta = \begin{bmatrix}\beta_d \\ \beta_n\end{bmatrix}$

$\therefore \hat{X}\beta = [X \;\; \alpha I_n]\begin{bmatrix}\beta_0 \\ \beta_n\end{bmatrix}$

$= X\beta_{01} + \alpha I_n \beta_n$

$= X\beta_d + \alpha\beta_n = y \qquad \therefore \beta_n = \frac{1}{\alpha}(y - X\beta_d)$

$\therefore \|\beta\|^2 = \|\beta_d\|^2 + \|\beta_n\|^2$

$= \|\beta_d\|^2 + \frac{1}{\alpha^2}\|y - X\beta_d\|^2$

$\therefore \frac{\partial\|\beta\|^2}{\partial\beta_d} = 2\beta_d + \frac{1}{\alpha^2}\cdot 2(-X^Ty + X^TX^T\beta_d)$

$= 2\beta_{d1} + \frac{2}{\alpha^2}(-X^Ty + X^TX\beta_d) = 0$

$\therefore \beta_d = (X^TX + \alpha^2 I_d)^{-1}X^Ty$

$\therefore \alpha = \sqrt{\lambda}$

$\therefore \beta_n = \frac{y - X(X^TX + \lambda I_d)^{-1}X^Ty}{\sqrt{\lambda}}$

(c) As $X' = XV$, $w' = V^Tw$

$\therefore \|y - X'w'\|_2^2 + \lambda\|w'\|_2^2 = \|y - XVV^Tw\|^2 + \lambda\|V^Tw\|^2$

As $X = U\Sigma V^T$ and V is identity

$\therefore \|y - X'w'\|_2^2 + \lambda\|w'\|_2^2 = \|y - Xw\|^2 + \lambda\|w\|^2$

$= \|y - U\Sigma V^Tw\|^2 + \lambda\|w\|^2$

$= y^Ty - 2y^TU\Sigma w + (U\Sigma w)^T\cdot U\Sigma w + \lambda w^Tw$

$$\therefore \frac{\partial(\|y - x'w'\|_2^2 + \lambda\|w'\|_2^2)}{\partial w} = -2(y^T \cdot U\Sigma)^T + 2\Sigma^T\Sigma w + 2\lambda w = 0$$

$$\therefore w = (\Sigma^T\Sigma + \lambda I)^{-1} \cdot (y^T U \Sigma V^T)^T$$

$$= (\Sigma^T\Sigma + \lambda I)^{-1} \cdot X^T y$$

As $\Sigma^T\Sigma + \lambda \cdot I = \begin{bmatrix} \sigma_1^2+\lambda & & \\ & \sigma_2^2+\lambda & \\ & & \ddots \\ & & & \sigma_d^2+\lambda \end{bmatrix}$

$$\therefore (\Sigma^T\Sigma + \lambda I)^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2+\lambda} & & & \\ & \frac{1}{\sigma_2^2+\lambda} & & \\ & & \ddots & \\ & & & \frac{1}{\sigma_d^2+\lambda} \end{bmatrix}$$

$$\therefore w_j = w_j = \sum_{i=1}^{n} \frac{1}{\sigma_i^2+\lambda} \cdot x_{ij}\, y_i$$

3. (a) ① $E[\hat{X}-\mu] = E[\frac{X_1+X_2+\cdots+X_n}{n}-\mu] = \frac{1}{n}E[X_1]+\frac{1}{n}E[X_2]+\cdots+\frac{1}{n}E[X_n]-\mu$

$\because E[X_i] = \mu$ for every i in $1,2,\cdots n$

$\therefore Bias[\hat{X}] = \frac{1}{n}\cdot n\cdot\mu-\mu = 0$

$Var[\hat{X}] = Var[\frac{X_1+X_2+\cdots+X_n}{n}] = \frac{1}{n^2}Var[X_1]+\frac{1}{n^2}Var[X_2]+\cdots+\frac{1}{n^2}Var[X_n]$

$\because Var[X_i] = \sigma^2$ for every i in $1,2,\cdots n$

$\therefore Var[\hat{X}] = \frac{1}{n^2}\cdot n\cdot\sigma^2 = \frac{\sigma^2}{n}$

② $E[\hat{X}-\mu] = E[\frac{X_1+X_2+\cdots+X_n}{n+n_0}-\mu]$

The same as above, $Bias[\hat{X}] = \frac{1}{n+n_0}\cdot n\cdot\mu-\mu = -\frac{n_0}{n+n_0}\mu$

$Var[\hat{X}] = Var[\frac{X_1+X_2+\cdots+X_n}{n+n_0}]$

The same as above, $Var[\hat{X}] = \frac{1}{(n+n_0)^2}\cdot n\cdot\sigma^2 = \frac{n\cdot\sigma^2}{(n+n_0)^2}$

(b) $E[(\hat{X}-\mu)^2] = E[(\hat{X}-E(\hat{X})+E(\hat{X})-\mu)^2]$

$= E[(\hat{X}-E(\hat{X}))^2]+2E[(\hat{X}-E(\hat{X}))(E(\hat{X})-\mu)]+E[(E(\hat{X})-\mu)^2]$

$= Var[\hat{X}]+2(E[\hat{X}]-E[E(\hat{X})])(E(\hat{X})-\mu)+E[(Bias[\hat{X}])^2]$

$= Var[\hat{X}]+(Bias[\hat{X}])^2$

$E[(\hat{X}-X')^2] = E[(\hat{X}-E(\hat{X})+E(\hat{X})-X')^2]$

$= Var[\hat{X}]+2E[(\hat{X}-E(\hat{X}))\cdot(E(\hat{X})-X')]+E[(E(\hat{X})-X')^2]$

$= Var[\hat{X}]+2E[\hat{X}E(\hat{X})-\hat{X}X'-E(\hat{X})^2+E(\hat{X})X']+E[E(\hat{X})^2+X'^2-2E(\hat{X})X']$

$= Var[\hat{X}]+2[E(\hat{X})^2-E(\hat{X}X')-E(\hat{X})^2+E(\hat{X})\cdot E(X')]+E(\hat{X})^2+E(X'^2)$

$= Var[\hat{X}]-2E(\hat{X}X')+E(\hat{X})^2+E(X'^2)$ $-2E(\hat{X})E(X')$

$= Var[\hat{X}]+Bias[\hat{X}]^2+2E[\hat{X}]\mu-\mu^2-2E[\hat{X}X']+E[X'^2]$

$= Var[\hat{X}]+Bias[\hat{X}]^2-2Cov(\hat{X},X')+Var[X']$

As $\hat{X}$ and X' are i.i.d, so $Cov(\hat{X},X') = 0$

$\therefore E[(\hat{X}-X')^2] = Var[\hat{X}]+Bias[\hat{X}]^2+Var[X']$

$E[(\hat{X}-X')^2]$ is bigger than $E[(\hat{X}-\mu)^2]$ and the difference between them is the variance of X'

(c). $E[(\hat{x}-\mu)^2] = Var[\hat{x}] + (Bias[\hat{x}])^2$

way 1: $E[(\hat{x}-\mu)^2] = \frac{\sigma^2}{n} + 0 = \frac{\sigma^2}{n}$

way 2: $E[(\hat{x}-\mu)^2] = \frac{n\sigma^2}{(n+n_0)^2} + \frac{n_0^2}{(n+n_0)^2}\mu^2$

$= \frac{n\sigma^2 + n_0^2\mu^2}{(n+n_0)^2}$.