CIS 520, Machine Learning, Fall 2015: Assignment 1

Arpit Panwar

September 11, 2015

Collaborators:

Anusha Fernando

1 High dimensional hi-jinx

1. Intra-class distance.

$$\begin{aligned}
\mathbf{E}[(X - X')^2] &= E[X^2 + X'^2 - 2 * X * X'] \\
&\quad Using\ \mathbf{E}[X + Y] = E[X] + E[Y]\ and\ \mathbf{E}[aX] = aE[X] \\
&= \mathbf{E}[X^2] + E[X'^2] + 2 * E[X * X'] \\
&\quad Replacing\ E[X^2] = \mu^2 + \sigma^2\ and\ E[X] = E[X'] = \mu \\
&= \mu^2 + \sigma^2 + \sigma^2 + \mu^2 - 2 * \mu^2 \\
&= 2 * \mu^2 + 2 * \sigma^2 - 2 * \mu^2 \\
&= 2 * \sigma^2
\end{aligned}$$

2. Inter-class distance.

$$\begin{aligned}
\mathbf{E}[(X - X')^2] &= E[X^2 + X'^2 - 2 * X * X'] \\
&\quad Using\ \mathbf{E}[X + Y] = E[X] + E[Y]\ and\ \mathbf{E}[aX] = aE[X] \\
&= \mathbf{E}[X^2] + E[X'^2] + 2 * E[X * X'] \\
&\quad Replacing\ E[X^2] = \mu_1^2 + \sigma^2\ , E[X'^2] = \mu_2^2 + \sigma^2,\ E[X] = \mu_1\ E[X'] = \mu_2 \\
&= \mu_1^2 + \sigma^2 + \sigma^2 + \mu_2^2 - 2 * \mu_1 * \mu_2 \\
&= 2 * \sigma^2 + (\mu_1 - \mu_2)^2
\end{aligned}$$

3. Intra-class distance, m-dimensions.

$$\mathbf{E}[\sum_{j=1}^{m}(X_j - X_j')^2] = \mathbf{E}[\sum_{j=1}^{m}(X_j^2 + {X_j'}^2 - 2 * X * X')]$$

$$Using\ \mathbf{E}[\sum_{j=1}^{m} X] = \sum_{j=1}^{m} E[X]$$

$$= \sum_{j=1}^{m}\mathbf{E}[X_j^2] + \sum_{j=1}^{m}\mathbf{E}[{X_j'}^2] - \sum_{j=1}^{m} 2 * \mathbf{E}[X_j] * \mathbf{E}[X_j']$$

$$Replacing\ E[X_j^2] = \mu_j^2 + \sigma^2\ and\ E[X_j] = E[X_j'] = \mu_j$$

$$= \sum_{j=1}^{m}(\mu_j^2 + \sigma^2 + \mu_j^2 + \sigma^2 - 2 * \mu_j * \mu_j)$$

$$= \sum_{j=1}^{m}(2 * \sigma^2)$$

$$= 2 * m * \sigma^2$$

4. Inter-class distance, m-dimensions.

$$\mathbf{E}[\sum_{j=1}^{m}(X_j - X_j')^2] = \mathbf{E}[\sum_{j=1}^{m}(X_j^2 + {X_j'}^2 - 2 * X * X')]$$

$$Using\ \mathbf{E}[\sum_{j=1}^{m} X] = \sum_{j=1}^{m} E[X]$$

$$= \sum_{j=1}^{m}\mathbf{E}[X_j^2] + \sum_{j=1}^{m}\mathbf{E}[{X_j'}^2] - \sum_{j=1}^{m} 2 * \mathbf{E}[X_j] * \mathbf{E}[X_j']$$

$$Replacing\ E[X_j^2] = \mu_1^2 + \sigma^2\ , E[{X_j'}^2] = \mu_2^2 + \sigma^2,\ E[X_j] = \mu_1\ E[X_j'] = \mu_2$$

$$= \sum_{j=1}^{m}(\mu_1^2 + \sigma^2 + \mu_2^2 + \sigma^2 - 2 * \mu_1 * \mu_2)$$

$$= \sum_{j=1}^{m}(2 * \sigma^2 + (\mu_1 - \mu_2)^2)$$

$$= \sum_{j=1}^{m}(2 * \sigma^2) + \sum_{j=1}^{m}(\mu_1 - \mu_2)^2$$

$$= 2 * m * \sigma^2 + \sum_{j=1}^{m}(\mu_1 - \mu_2)^2$$

5. The ratio of expected intra-class distance to inter-class distance is: $\frac{2 * m * \sigma^2}{2 * m * \sigma^2 + (\mu_1 - \mu_2)^2}$.

   As m increases towards ∞, this ratio approaches

$$\lim_{m \to \infty} \frac{2 * m * \sigma^2}{2 * m * \sigma^2 + (\mu_1 - \mu_2)^2}$$

$$Using L'Hopital'srule$$

$$= \frac{2 * \sigma^2}{2 * \sigma^2}$$

$$= 1$$

2 Fitting distributions with KL divergence

1. KL divergence for Gaussians.

   (a) The KL divergence between two univariate Gaussians is given by $f = \ldots$ and $g = \ldots$.

$$
\begin{aligned}
KL(p(x)||q(x)) &= \mathbf{E}_p\left[\log_e \frac{p(x)}{q(x)}\right] \\
&= \mathbf{E}_p\left[log_e \frac{\frac{1}{\sqrt{2*\pi*\sigma}} * e^{-\dfrac{(x-\mu_1)^2}{2*\sigma^2}}}{\sqrt{2*\pi} * e^{-\dfrac{(x-\mu_2)^2}{2}}}\right] \\
&= \mathbf{E}_p\left[log_e \frac{1}{\sigma} * e^{\dfrac{(x-\mu_2)^2}{2} - \dfrac{(x-\mu_1)^2}{2*\sigma^2}}\right] \\
&= \mathbf{E}_p\left[log_e \frac{1}{\sigma} + log_e e^{\dfrac{(x-\mu_2)^2}{2} - \dfrac{(x-\mu_1)^2}{2*\sigma^2}}\right] \\
&= \mathbf{E}_p\left[\frac{(x-\mu_2)^2}{2} - \frac{(x-\mu_1)^2}{2*\sigma^2}\right] + log_e \frac{1}{\sigma} \\
&= \mathbf{E}_p[f(x, \mu_1, \mu_2, \sigma)] + g(\sigma)
\end{aligned}
$$

(b) The value $\mu_1 = \ldots$ minimizes $KL(p(x)||q(x))$.

$$KL(p(x)||q(x)) = \int_{-\infty}^{\infty} p(x) * log_e \left(\frac{\frac{1}{\sqrt{2*\pi*\sigma}} * e^{-\dfrac{(x-\mu_1)^2}{2*\sigma^2}}}{\frac{1}{\sqrt{2*\pi}} * e^{-\dfrac{(x-\mu_2)^2}{2}}} \right)$$

$$= \int_{-\infty}^{\infty} p(x) * log_e \left(\frac{1}{\sigma} * e^{\dfrac{(x-\mu_2)^2}{2} - \dfrac{(x-\mu_1)^2}{2*\mu^2}} \right)$$

$$Expanding\ logarithm\ and\ solving\ log_e(e^x)$$

$$= \int_{-\infty}^{\infty} p(x) * log_e \left(\frac{1}{\sigma} \right) + \int_{-\infty}^{\infty} p(x) * \frac{(x-\mu_2)^2}{2} - \int_{-\infty}^{\infty} p(x) * \frac{(x-\mu_1)^2}{2*\sigma^2}$$

$$We\ know\ that \int p(x) * (x-\mu)^2 = \sigma^2\ and \int p(x)dx = 1 \ldots\ldots\ldots (i)$$

$$Substituting\ the\ values\ in\ above\ equation$$

$$= log_e \left(\frac{1}{\sigma} \right) + \int_{-\infty}^{\infty} p(x)(x-\mu_1+\mu_1-\mu_2)^2 dx - \int_{-\infty}^{\infty} \frac{1}{2} dx$$

$$= log_e \left(\frac{1}{\sigma} \right) + \int_{-\infty}^{\infty} p(x) \left[(x-\mu_1)^2 + (\mu_1-\mu_2)^2 + 2*(x-\mu_1)*(\mu_1-\mu_2) \right] - \frac{1}{2}$$

$$= log_e \left(\frac{1}{\sigma} \right) - \frac{1}{2} + \int_{-\infty}^{\infty} p(x)*(x-\mu_1)^2 + \int_{-\infty}^{\infty} p(x)(\mu_1-\mu_2)^2 + \int_{-\infty}^{\infty} p(x)*2*(x-\mu_1)*(\mu_1-\mu_2)$$

$$Again\ using (i)$$

$$= log_e \left(\frac{1}{\sigma} \right) - \frac{1}{2} + \sigma^2 + (\mu_1-\mu_2)^2 + (\mu_1-\mu_2) * \left[\left(\int_{-\infty}^{\infty} p(x)*2*x\ dx \right) - \left(\mu_1 * \int_{-\infty}^{\infty} p(x)dx \right) \right]$$

$$= log_e \left(\frac{1}{\sigma} \right) - \frac{1}{2} + \sigma^2 + (\mu_1-\mu_2)^2 + (\mu_1-\mu_2)*0$$

$$= log_e \left(\frac{1}{\sigma} \right) - \frac{1}{2} + \sigma^2 + (\mu_1-\mu_2)^2$$

$$0 = \frac{\partial KL(p(x)||q(x))}{\partial \mu_1}$$

$$0 = \frac{\partial}{\partial \mu_1} log_e \left(\frac{1}{\sigma} \right) - \frac{1}{2} + \sigma^2 + (\mu_1-\mu_2)^2$$

$$0 = 2*\mu_1 - 2*\mu_2$$

$$\mu_1 = \mu_2$$

2. KL divergence for Multinomials.

(a) The KL divergence between two Multinomials is: $KL(p(x)||q(x)) = \sum_i p(x) * log\left(\frac{p(x)}{q(x)}\right)$.

$$\begin{aligned}
KL(p(x)||q(x)) = & \sum_{i\ even} p(x_i) log_e\left(\frac{p(x_i)}{q(x_i)}\right) + \sum_{i\ odd} p(x_i) log_e\left(\frac{p(x_i)}{q(x_i)}\right) \\
& Substituting\ the\ values \\
= & \sum_{i\ even} \alpha \log_e\left(\frac{\alpha}{\theta_{even}}\right) + \sum_{i\ odd} \beta \log_e\left(\frac{\beta}{\theta_{odd}}\right) \\
& Solving\ logarithms \\
= & \sum_{i\ even} \alpha \log_e(\alpha) - \sum_{i\ even} \alpha \log_e(\theta even) + \sum_{i\ odd} \beta \log_e(\beta) - \sum_{i\ odd} \beta \log_e(\theta_{odd}) \\
= & n * \alpha * \log_e(\alpha) + n * \beta * \log_e(\beta) - \sum_{i\ even} \alpha \log_e(\theta even) - \sum_{i\ odd} \beta \log_e(\theta_{odd}) \ldots\ldots\ldots (i)
\end{aligned}$$

(b) The values $\alpha = \ldots$ and $\beta = \ldots$ minimize $KL(p(x)||q(x))$. We are given that to find the minimum we need to add the Lagranges multiplier and set to 0. We need to minimize $KL(p(x)||q(x)) + \lambda(n(\alpha + \beta) - 1)$

Substituting values in (i)

$$n*\alpha*\log_e(\alpha)+n*\beta*\log_e(\beta)-\sum_{i\ even}\alpha\log_e(\theta_{even})-\sum_{i\ odd}\beta\log_e(\theta_{odd})+\lambda(n(\alpha+\beta)-1)=0$$

$$\frac{\partial}{\partial\alpha}=n*\log_e(\alpha)+n+n\lambda-\sum_{i\ even}\log_e(\theta_{even})=0\ldots\ldots\ldots(ii)$$

$$\frac{\partial}{\partial\beta}=n*\log_e(\beta)+n+n\lambda-\sum_{i\ odd}\log_e(\theta_{odd})=0\ldots\ldots\ldots(iii)$$

$$\frac{\partial}{\partial\lambda}=n\alpha+n\beta-1=0\ldots\ldots\ldots(iv)$$

Adding (ii) and (iii)

$$n\log_e(\alpha)+n\log(\beta)+2n(\lambda+1)-\sum_{i\ odd}\log_e(\theta_{odd})-\sum_{i\ even}\log_e(\theta_{even})=0\ldots\ldots\ldots(v)$$

Subtracting (iii) from (ii)

$$n\log(\alpha)-n\log(\beta)+\sum_{i\ odd}\log_e(\theta_{odd})-\sum_{i\ even}\log_e(\theta_{even})=0\ldots\ldots\ldots(vi)$$

Adding (v) and (vi)

$$2*n*\log(\alpha)+2*n*(\lambda+1)=2*\sum_{i\ even}\log_e(\theta_{even})$$

$$\frac{\sum_{i\ even}\log_e(\theta_{even})}{n}-(\lambda+1)=\log_e(\alpha)$$

$$e^{\frac{\sum_{i\ even}\log_e(\theta_{even})}{n}-(\lambda+1)}=\alpha\ldots\ldots\ldots(vii)$$

Subtracting (vi) from (v)

$$2*n*\log_e\beta+2*n*(\lambda+1)=2\sum_{i\ odd}\log_e(\theta_{odd})$$

$$\frac{\sum_{i\ odd}\log_e(\theta_{odd})}{n}-(\lambda+1)=\log_e(\beta)$$

$$e^{\frac{\sum_{i\ odd}\log_e(\theta_{odd})}{n}-(\lambda+1)}=\beta\ldots\ldots\ldots(viii)$$

Substituting value of (vii)and (viii) in (iv)

$$e^{\frac{\sum_{i\ even}\log_e(\theta_{even})}{n}-(\lambda+1)}+e^{\frac{\sum_{i\ odd}\log_e(\theta_{odd})}{n}-(\lambda+1)}=\frac{1}{n}$$

$$e^{\frac{\sum_{i\ even}\log_e(\theta_{even})}{n}}+e^{\frac{\sum_{i\ odd}\log_e(\theta_{odd})}{n}}=\frac{e^{(\lambda+1)}}{n}\ldots\ldots\ldots(ix)$$

Solving for λ we get

$$\log n+\log_e\left(e^{\frac{\sum_{i\ even}\log_e(\theta_{even})}{n}}+e^{\frac{\sum_{i\ odd}\log_e(\theta_{odd})}{n}}\right)-1=\lambda$$

Substituting (ix) in (vii)

$$\frac{e^{\frac{\sum_{i\ even}\log_e(\theta_{even})}{n}}}{n*e^{\frac{\sum_{i\ even}\log_e(\theta_{even})}{n}}+e^{\frac{\sum_{i\ odd}\log_e(\theta_{odd})}{n}}}=\alpha$$

Substituting (ix) in (viii)

$$\frac{6\qquad e^{\frac{\sum_{i\ odd}\log_e(\theta_{odd})}{n}}}{n*e^{\frac{\sum_{i\ even}\log_e(\theta_{even})}{n}}+e^{\frac{\sum_{i\ odd}\log_e(\theta_{odd})}{n}}}=\beta$$

3 Conditional independence in probability models

1. We can write

$$\begin{aligned}
&UsingMarginalization\\
p(x_i) &= p(x_i|z_i = 1) * p(z_i = 1) + \ldots + p(x_i|z_i = j) * p(z_i = j) + \ldots + p(x_i|z_i = k) * p(z_i = k)\\
&= \sum_{j=1}^{k} p(x_i|z_i = j)p(z_i = j)\\
&= \sum_{j=1}^{k} f_j(x_i)\pi_j
\end{aligned}$$

2. The formula for $p(x_1, \ldots, x_n)$ is ...

$$\begin{aligned}
&\textit{Assuming all the } x_i \textit{ are independent}\\
p(x_1, \ldots, x_n) &= p(x_1) * p(x_2) * \ldots * p(x_n)\\
&\textit{Substituting } p(x_i) \textit{ from part i}\\
&= \sum_{j=1}^{k} f_j(x_1)\pi_j * \sum_{j=1}^{k} f_j(x_2)\pi_j * \ldots * \sum_{j=1}^{k} f_j(x_n)\pi_j\\
&= \prod_{i=1}^{n}\sum_{j=1}^{k} f_j(x_i)\pi_j
\end{aligned}$$

3. The formula for $p(z_u = v \mid x_1, \ldots, x_n)$ is ...

$$\begin{aligned}
p(z_u = v \mid x_1, \ldots, x_n) &= \frac{p(x_1 \ldots \mid z_u = v) * p(z_u = v)}{p(x_1, x_2 \ldots x_n)}\\
&= \frac{p(x_1, x_2 \ldots x_{u-1}, x_{u+1} \ldots x_n) * p(x_u \mid z_u = v) * p(z_u = v)}{p(x_1, x_2 \ldots x_n)}\\
&= \frac{p(x_u \mid z_u = v) * p(z_u = v)}{p(x_u)}\\
&= \frac{f_v(x_u) * \pi_v}{\sum_{j=1}^{k} f_j(x_u) * \pi_j}
\end{aligned}$$

4 Decision trees

1. Concrete sample training data.

   (a) The sample entropy $H(Y)$ is ....

$$\begin{aligned}
H(Y) &= -\sum_{y} P(Y = y) \log_2 P(Y = y)\\
&= -[3/5 * log_2(3/5)] - [2/5 * log_2(2/5)]\\
&= 0.4422 + 0.5288\\
&= 0.9710
\end{aligned}$$

(b) The information gains are $IG(X_1) = \ldots$ and $IG(X_2) = \ldots$.

$$\begin{aligned}
IG(X_1) &= H(Y) - H(Y \mid X_1) \\
&\quad \textit{We already have } H(Y) \textit{ from part 1 above. So calculating } H(Y \mid X_1) \\
H(Y \mid X_1) &= H(Y \mid x_1 = T) * P(X_1 = T) + H(Y \mid x_1 = F) * P(X_1 = F) \\
&= -[(2/3 * \log_2(2/3) + 1/3 * \log_2(1/3)) * 9/20] - [(6/11 * \log_2(6/11) + (5/11) * \log_2(5/11)) * 11/20] \\
&= -0.4132 - 0.5467 \\
&= -0.9599 \\
IG(X_1) &= 0.9710 - 0.9599 = 0.0111 \\
IG(X_2) &= H(Y) - H(Y \mid X_2) \\
&\quad \textit{We already have } H(Y) \textit{ from part 1 above. So calculating } H(Y \mid X_2) \\
H(Y \mid X_2) &= H(Y \mid x_2 = T) * P(X_2 = T) + H(Y \mid x_2 = F) * P(X_2 = F) \\
&= -[(4/5 * \log_2(4/5) + (1/5 * \log_2(1/5))) * 1/2] - [(2/5 * \log_2(2/5) + (3/5 * \log_2(3/5))) * 1/2] \\
&= -0.3610 - 0.4855 = -0.8465 \\
IG(X_2) &= 0.9710 - 0.8465 = 0.1245
\end{aligned}$$

(c) The decision tree that would be learned is shown in Figure 1.

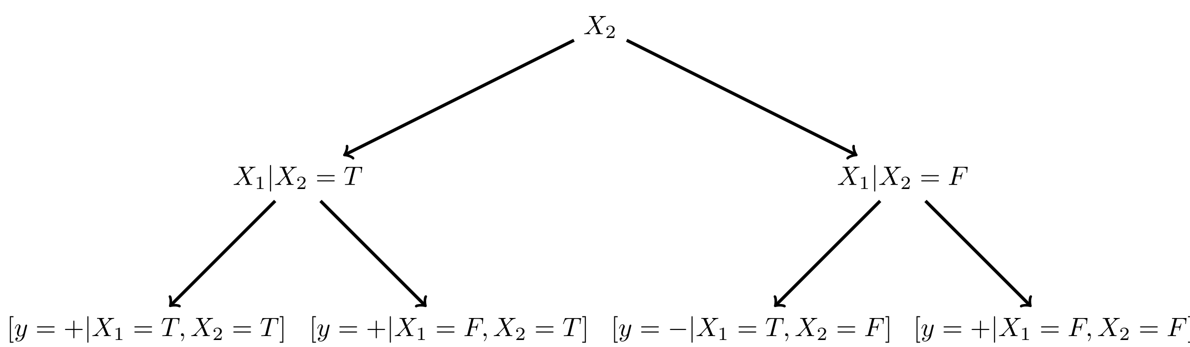

Figure 1: The decision tree that would be learned.

From left to right below is the reason for choosing the sign of y
i. y=+ : count + = 6 , count - = 1
ii. y=+ : count + = 2 , count - = 1
iii. y=- : count + =0, count - = 2
iv. y=+ : count + =4, count -= 4

2. Proof that $IG(x, y) = H[x] - H[x \mid y] = H[y] - H[y \mid x]$, starting from the definition in terms of

KL-divergence:

$$
\begin{aligned}
IG(x,y) &= KL\,(p(x,y)||p(x)p(y)) \\
&= -\sum_x \sum_y p(x,y) * \log_2(\frac{p(x)*p(y)}{p(x,y)}) \\
&= \sum_x \sum_y p(x,y) * \log_2(\frac{p(x,y)}{p(x)*p(y)}) \ldots\ldots (i) \\
&\quad Using\ bayes\ rule \\
&= \sum_x \sum_y p(x,y) * \log_2(\frac{p(x|y)}{p(x)}) \\
&\quad Splitting\ logarithm \\
&= \sum_x \sum_y p(x,y) * \log_2(p(x|y)) - \sum_x \sum_y p(x,y) * \log_2(p(x)) \\
&\quad Replacing\ with\ the\ definition\ of\ Entropy\ and\ Marginalize\ right\ hand\ term\ over\ y \\
&= -H[x \mid y] - \sum_x \log_2(p(x))[p(x_i,y_1) + p(x_i,y_2) + \ldots + p(x_i,y_n)] \\
&= -H[x \mid y] - \sum_x \log_2(p(x)) * p(x) \\
&= -H[x \mid y] + H[x] \\
&= H[x] - H[x \mid y] \\
\\
&\quad Again\ using\ (i)\ and\ applying\ bayes\ rule\ again \\
\\
&= \sum_x \sum_y p(x,y) * \log_2(\frac{p(y|x)}{p(y)}) \\
&\quad Splitting\ logarithm \\
&= \sum_x \sum_y p(x,y) * \log_2(p(y|x)) - \sum_x \sum_y p(x,y) * \log_2(p(y)) \\
&\quad Replacing\ with\ the\ definition\ of\ Entropy\ and\ Marginalize\ right\ hand\ term\ over\ x \\
&= -H[y \mid x] - \sum_y \log_2(p(x))[p(x_1,y_i) + p(x_1,y_i) + \ldots + p(x_n,y_i)] \\
&= -H[y \mid x] - \sum_x \log_2(p(x)) * p(x) \\
&= -H[y \mid x] + H[y] \\
&= H[y] - H[y \mid x]
\end{aligned}
$$

5 Non-Normal Norms

1. For the given vectors,the point closest to x_1 under each of the following norms is

a) L_0:

$$\begin{aligned}
&\textit{Calculating } L_0 \textit{ distance}\\
&x_1 - x_2 = [0.4, -1.5, -1.6, 0.9]\\
&x_1 - x_3 = [0, 0, -0.7, -7.1]\\
&x_1 - x_4 = [0.9, 0.5, -2, -0.5]
\end{aligned}$$

So. x_3 is the closest with distance [0,0,0.7,7.1]

b) L_1:

$$\begin{aligned}
&\textit{Calculating } L_1 \textit{ distance}\\
&\sum[x_1 - x_2] = [4.4]\\
&\sum[x_1 - x_3] = [7.8]\\
&\sum[x_1 - x_4] = [3.9]
\end{aligned}$$

So, x_4 is the closest with distance 3.9 c) L_2:

$$\sqrt{\sum_i (x_{1i} - x_{2i})^2} = 2.3748$$

$$\sqrt{\sum_i (x_{1i} - x_{3i})^2} = 7.1344$$

$$\sqrt{\sum_i (x_{1i} - x_{4i})^2} = 2.3043$$

So, x_4 is closest with distance 2.3043

d) L_{inf}:

$$\begin{aligned}
\max(x_1 - x_2) &= 1.6\\
\max(x_1 - x_3) &= 7.1\\
\max(x_1 - x_4) &= 2
\end{aligned}$$

So x_2 is the closest with distance 1.6

2. i) "loss function" of $y - \hat{y} =$

$$f(y - \hat{y}) = \begin{cases} k, & \text{if } (y - \hat{y}) > 0. \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

ii) By third rule for norm $p(x) = 0$ means x is a zero vector but in our case $p(x)$ is 0 even when x is non zero thus it is not a norm

iii) L_1 norm describes the error in the best possible form as for any fixed value of y we can fetch the of error function as a straight line distance between origin (value 0) and the value of $\hat{y}$