TD 1 Linear Regression:

MLE & MAP

0: Maximum Likehood Estimation (MLE)

- Guillame and Vassilis would like to know what percentage of students like the Introductory course to Machine Learning
- Let this unknown, but hopefully very close to 1, quantity be denoted by μ
- To estimate μ, the instructors created an anonymous survey which contains this question: "Do you like the ML course? Yes or No"
- Each student can only answer this question once, and we assume that the distribution of the answers is i.i.d.

(a) What is the MLE estimation of μ?

(b) Let the above estimator be denoted by $\hat{\mu}$. How many students should the instructors ask if they want the estimated value $\hat{\mu}$ to be so close to the unknown μ such that

$$\mathbb{P}(|\hat{\mu} - \mu| > 0.1) < 0.05$$

0: Solution

(a) This problem is equivalent to estimating the mean parameter of a Bernoulli distribution from i.i.d. data

Parameter	$p \in [0,1]$: success probability
Support	$\{0,1\}$
PMF	$p^x(1-p)^{1-x}$
Mean	p
Variance	$pq = p(1-p)$
MGF	$(1-p) + pe^t, \qquad (t \in \mathbb{R})$

- Therefore, the MLE estimation is $\mu\hat{} = n_1/N$, where n_1 is the number of students who answered Yes and N is the total number of students

0: Solution

(b) Recall Hoeffding's inequality (1963) providing an upper bound on the probability that the sum of bounded independent random variables deviates from its expected value by more than a certain amount

$$\left.\begin{array}{r} X_1, \ldots, X_n \text{ independent} \\ X_i \in [a_i, b_i] \\ \varepsilon > 0 \end{array}\right\} \Rightarrow$$

$$\Rightarrow \mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)| > \varepsilon) \leq 2\exp\left(\frac{-2n\varepsilon^2}{\frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

- It only contains the range $[a_i , b_i]$ of the random variables X_i is the range of the variable, but not the variances !

0: Solution

- Let X_i = 1 if a student answered yes, and let X_i = 0 if the answer was no

- According to Hoeffding's equality,

$$\Pr(|\hat{\mu} - \mu| > \epsilon) \leq 2\exp\left(-\frac{2N^2\epsilon^2}{\sum_{i=1}^{N}(b_i - a_i)^2}\right)$$

[a_i , b_i] is the range of the random variable X_i, therefore in our case a_i = 0, b_i = 1

$$P(|\hat{\mu} - \mu| > 0.1) < 2e^{-2N\times(0.1)^2} = 0.05$$

from which we have, N = 50 ln40

- So, the instructors need 185 students

0 Bonus Exercise: Variance and Concentration (41 pts)

- Guillame and Vassilis would like to see if the students attending the ML course like probability theory
- You know (because you're so friendly) that 200 out of the 250 students in the ML course say they like probability theory, but instructors don't believe you
- They decide to use the following process to estimate the number of people who like probability theory:
- Choose a student uniformly at random and independent from any previous choices
  - $X_i = 1$ if the student likes probability
  - Record $X_i = 0$ otherwise
- They will choose 30 such students this way, and they define the average of X_i

$$X = \frac{\sum_{i=1}^{30} X_i}{30},$$

0 Bonus Exercise: Variance and Concentration (41 pts)

(a) **(4 points)** What is $E[X_1]$?

(b) **(4 points)** What is Var(X1)?

- Hint: p(1 - p) is the variance of a Bernoulli random variable with probability of success p

(c) **(6 points)** What is E[X]?

(d) **(7 points)** What is Var(X)?

(e) **(20 points)** Guillaume and Vassilis are worried that less than half the course likes probability theory: they will stop being worried if $X \geq 0.5$

- Hint: Use Chebyshev's inequality to give an upper bound on the probability that they should stop worrying

- **Chebyshev's Inequality**: If X is a random variable with finite mean μ and finite variance σ^2, then for any real number $k > 0$: $Pr\left[|X-\mu| \geq k\sigma\right] \leq \frac{1}{k^2}$

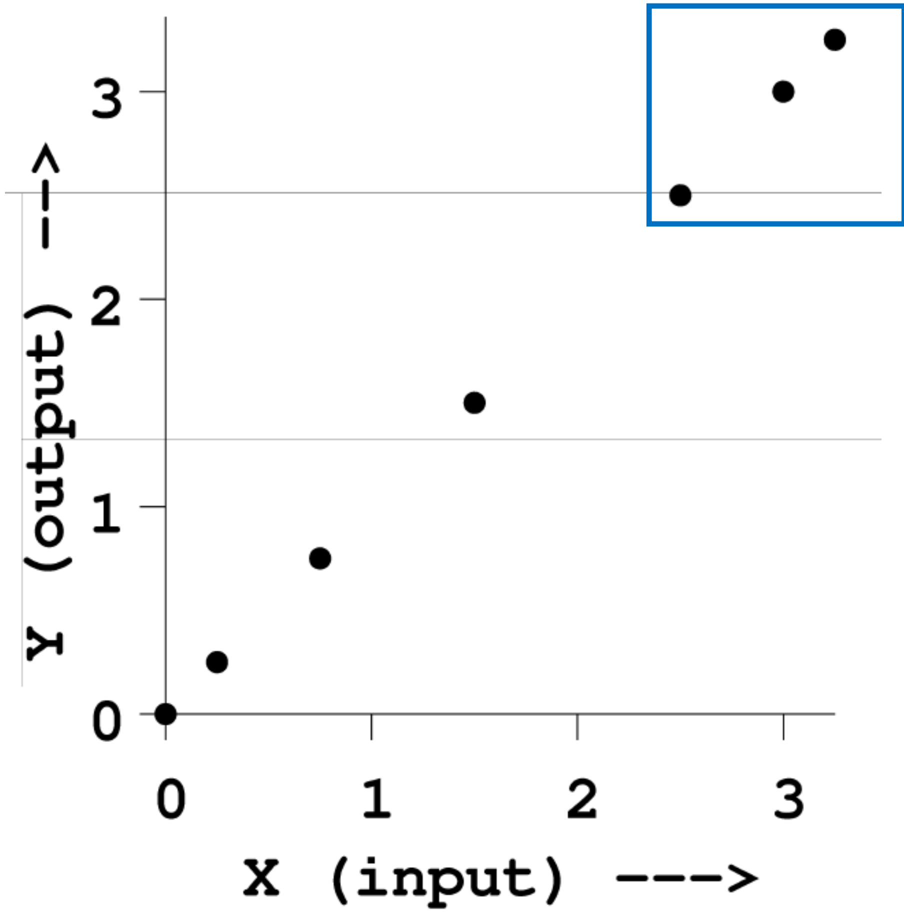
1 Train and Test Error

(a) Consider the following data with one input and one output

i) What is the mean squared training set error of running linear regression (using the model $y = w_0 + w_1 x$) on this data?

ii) What is the mean squared test set error of running linear regression on this data, assuming that the rightmost three points are in the test set, and the other are in the training set

iii) What is the mean squared leave-one out cross-validation (LOOCV) error of running linear regression on this data?



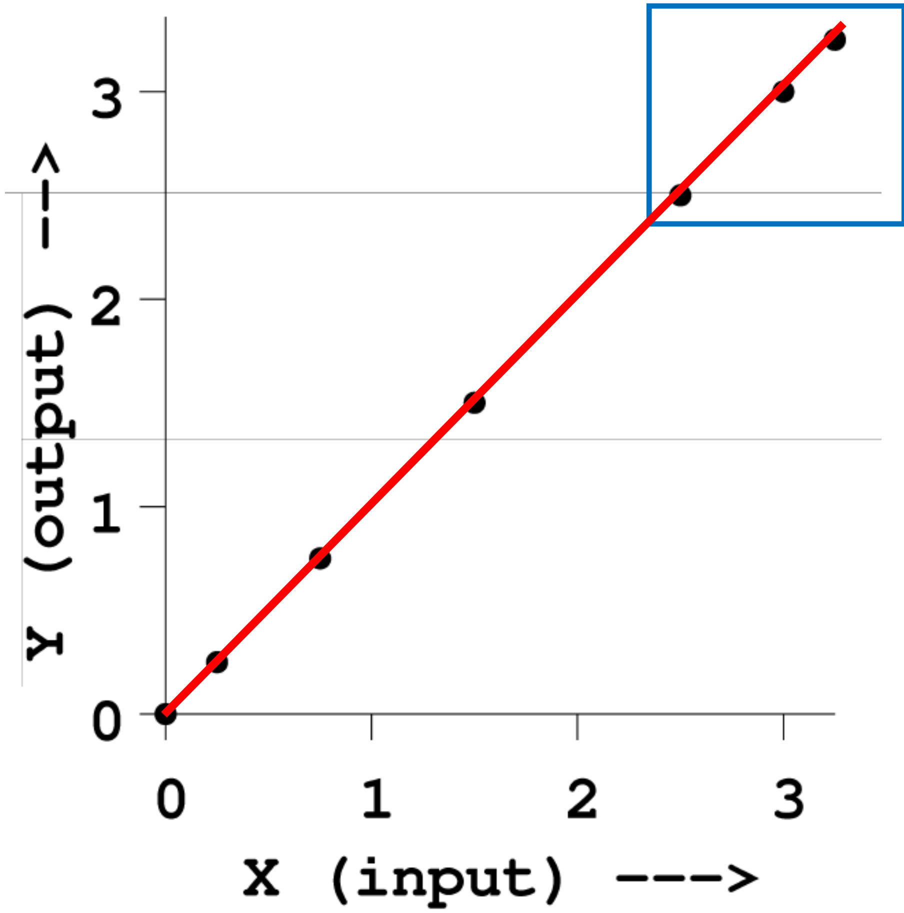
1 Solution

(a) Consider the following data with one input and one output

i) 0

ii) 0

iii) 0



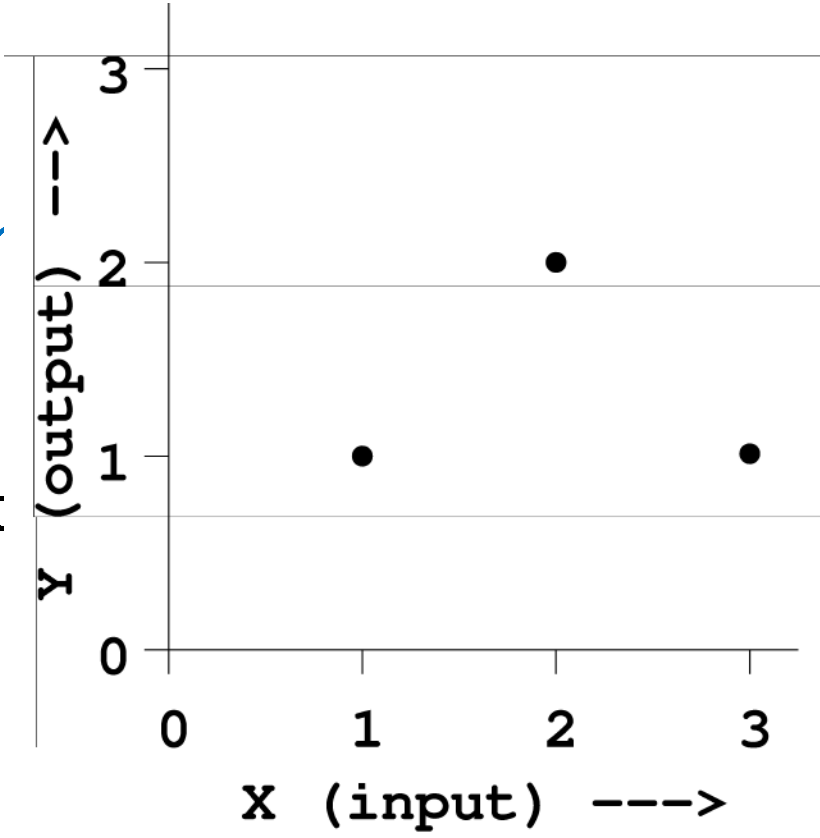
1 Train and Test Error

(b) Consider the following data with one input and one output

i) What is the mean squared training set error MSE of running linear regression (using the model $y = w_0 + w_1 x$) on this data?

- Hint: By symmetry we can see that the best fit to these three points is a horizontal line

ii) What is mean squared leave-one out cross-validation (LOOCV) error of running linear regression on this data?



X	Y
1	1
2	2
3	1

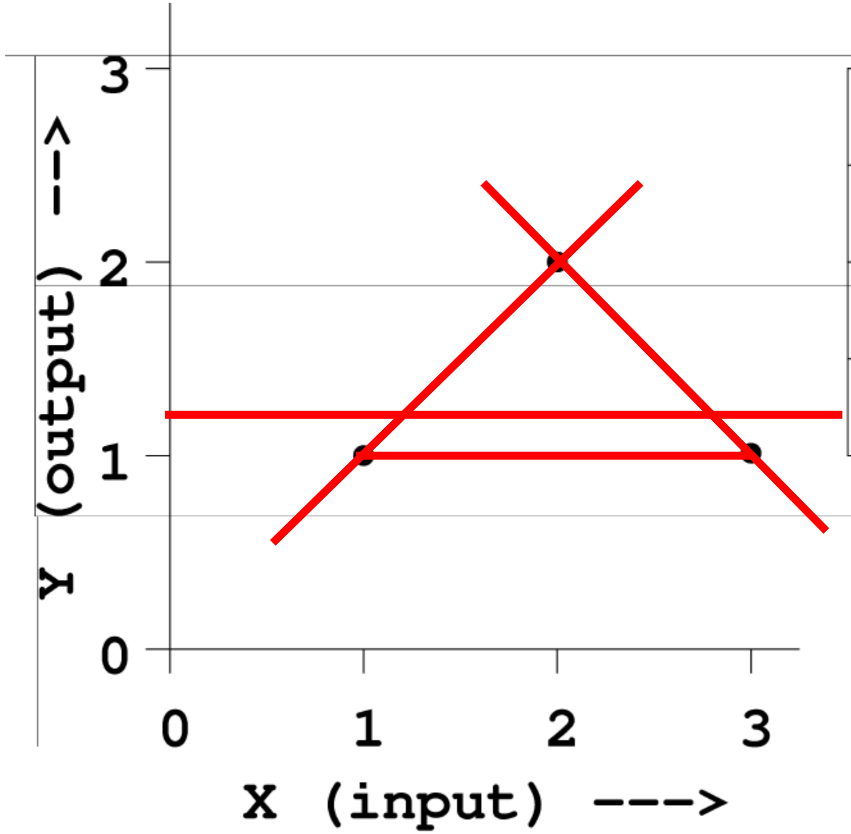
1 Solution

(a) Consider the following data with one input and one output

i) $SSE = (1/3)^2 + (1/3)^2 + (2/3)^2 = 6/9$

$MSE = SSE/3 = 2/9$

ii) $MSE = (1^2 + 1^2 + 1^2)/3 = 3$



X	Y
1	1
2	2
3	1

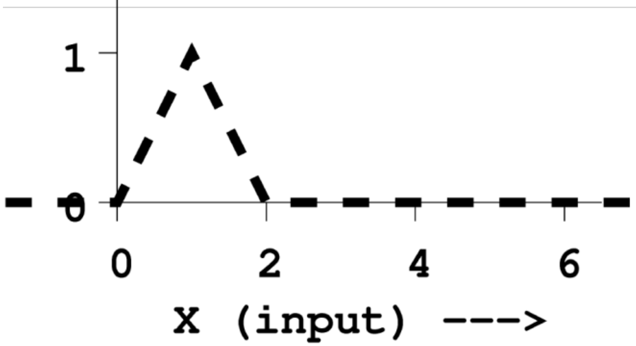
1 Train and Test Error

(c) Suppose that we plan multiple regression using the model $y = \beta_1\phi_1(x) + \beta_2\phi_2(x) + \beta_3\phi_3(x)$ with the following basic functions

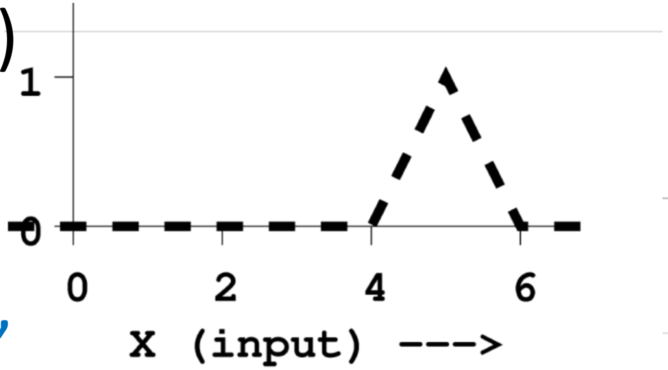
- Assume that all our training data points, and future queries lies between 1 and 5: $1 \leq X \leq 5$

(i) Is this a generally useful set of basic functions to use?

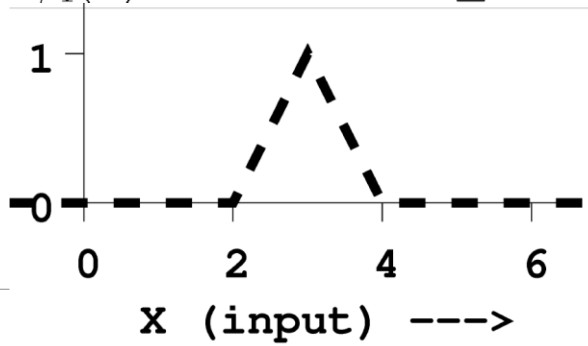
- If ‘yes’ explain their primer advantage
- If ‘no’ explain their biggest drawback



$$
\begin{array}{llll}
\phi_1(x) = & 0 & \text{if } x < 0 \\
\phi_1(x) = & x & \text{if } 0 \leq x < 1 \\
\phi_1(x) = & 2 - x & \text{if } 1 \leq x < 2 \\
\phi_1(x) = & 0 & \text{if } 2 \leq x
\end{array}
$$



$$
\begin{array}{llll}
\phi_2(x) = & 0 & \text{if } x < 2 \\
\phi_2(x) = & x - 2 & \text{if } 2 \leq x < 3 \\
\phi_2(x) = & 4 - x & \text{if } 3 \leq x < 4 \\
\phi_2(x) = & 0 & \text{if } 4 \leq x
\end{array}
$$



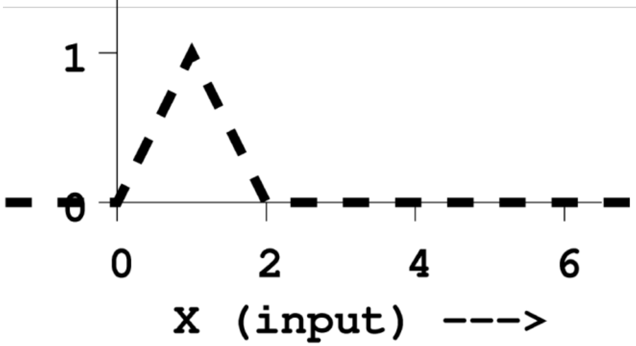
$$
\begin{array}{llll}
\phi_3(x) = & 0 & \text{if } x < 4 \\
\phi_3(x) = & x - 4 & \text{if } 4 \leq x < 5 \\
\phi_3(x) = & 6 - x & \text{if } 5 \leq x < 6 \\
\phi_3(x) = & 0 & \text{if } 6 \leq x
\end{array}
$$

1 Solution

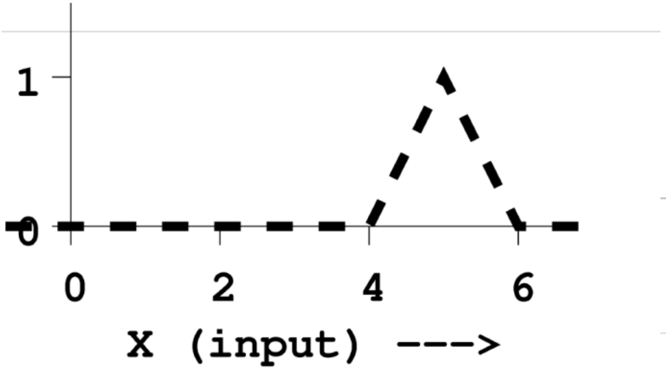
(c) Suppose that we plan multiple regression using the model $y = \beta_1\phi_1(x) + \beta_2\phi_2(x) + \beta_3\phi_3(x)$ with the following basic functions

i) No

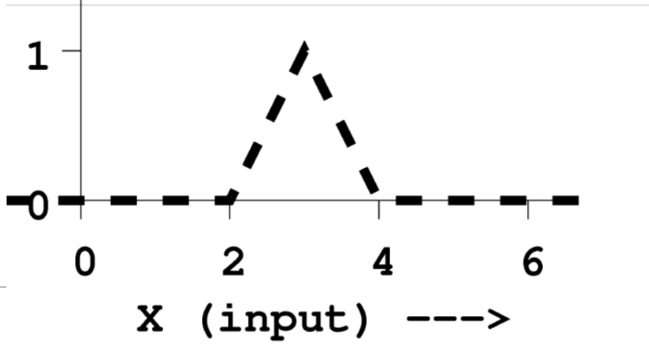
They are forced to predict y= 0 at X=2 and X=4 (and forced to be close to zero nearby) no matter what are the values of β



$$\begin{aligned}
\phi_1(x) &= 0 && \text{if } x < 0 \\
\phi_1(x) &= x && \text{if } 0 \le x < 1 \\
\phi_1(x) &= 2 - x && \text{if } 1 \le x < 2 \\
\phi_1(x) &= 0 && \text{if } 2 \le x
\end{aligned}$$



$$\begin{aligned}
\phi_3(x) &= 0 && \text{if } x < 4 \\
\phi_3(x) &= x - 4 && \text{if } 4 \le x < 5 \\
\phi_3(x) &= 6 - x && \text{if } 5 \le x < 6 \\
\phi_3(x) &= 0 && \text{if } 6 \le x
\end{aligned}$$



$$\begin{aligned}
\phi_2(x) &= 0 && \text{if } x < 2 \\
\phi_2(x) &= x - 2 && \text{if } 2 \le x < 3 \\
\phi_2(x) &= 4 - x && \text{if } 3 \le x < 4 \\
\phi_2(x) &= 0 && \text{if } 4 \le x
\end{aligned}$$

2: Optimal Mean Square Error (MSE) Rule

- Suppose we knew the joint distribution P(X,Y)
- The optimal rule $f^*: X \to Y$ which minimizes the MSE is given as:

$$f^* = \arg\min_f \mathbb{E}[(f(X) - Y)^2]$$

- Show that $f^*(X) = E[Y \mid X]$

- Hint: it suffices to argue that

$$\mathbb{E}_{X,Y}[(f(X) - Y)^2] \geq \mathbb{E}[(\mathbb{E}[Y|X] - Y)^2] \quad \text{for all } f$$

and hence $f^*(X) = E[Y \mid X]$

Properties of Conditional Expectation

- Let X, Y be discrete random variables: the conditional expectation $E[X|Y=y]$ can be seen as a function of random outcome ω: $\omega \rightarrow E[X|Y = Y(\omega)]$
- Theorem: Let X, Y, Z be random variables, $a, b \in R$, and $g : R \rightarrow R$. Assuming all the following expectations exist, we have

  (i) constant: $E[a|\ Y] = a$

  (ii) linearity: $E[aX + bZ|\ Y] = aE[X|\ Y] + bE[Z|\ Y]$

  (iii) Independence: $E[X|\ Y] = E[X]$ if X and Y are independent

  (iv) Adam's Law / Law of Iterated Expectation: $E[E[X|\ Y]] = E[X]$

  (v) Taking out what is known: $E[X f(Y)|\ Y] = f(Y) E[X|\ Y]$ and $E[X|\ Y f(Y)] = E[X|\ Y]$

  - In particular, $E[f(Y)|\ Y] = f(Y)$

  (vi) Keeping just what is needed: $E[X Y] = E[X E[Y | X]]$

  (vii) Projection interpretation: $E[(Y - E[Y | X]) f(X)] = 0$ for any function $f : X \rightarrow R$

  (viii) $E[\ E[X|Y, Z] \mid Y] = E[X|Y]$

2 Solution

$$\mathbb{E}[(f(X) - Y)^2] = \mathbb{E}[(\overbrace{f(X) - \mathbb{E}[Y|X]}^{a} + \overbrace{\mathbb{E}[Y|X] - Y}^{b})^2]$$

$$= \mathbb{E}[\overbrace{(f(X) - \mathbb{E}[Y|X])^2}^{a^2} + \overbrace{(\mathbb{E}[Y|X] - Y)^2}^{b^2} + \overbrace{2(f(X) - \mathbb{E}[Y|X])(\mathbb{E}(Y|X) - Y)}^{2ab}]$$

$$= \mathbb{E}[(f(X) - \mathbb{E}[Y|X])^2] + \mathbb{E}[(\mathbb{E}[Y|X] - Y)^2] + 2\mathbb{E}[(f(X) - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - Y)]$$

2 Solution

- Before knowing the realization of X, the conditional expectation of Y given X is unknown and can itself be regarded as a random variable E[Y |X]
  - In other words, E[Y |X] is a random variable such that its realization equals E[Y |X = x] when x is the realization of X
- Now using the law of iterated expectations (or tower property),

$\mathbb{E}_{XY}[\ldots] = \mathbb{E}_X[\mathbb{E}_{Y|X}[\ldots|X]]$ we have

$$\mathbb{E}_{XY}[(f(X)-\mathbb{E}[Y|X])(\mathbb{E}[Y|X]-Y)] = \mathbb{E}_X[\mathbb{E}_{Y|X}[(f(X)-\mathbb{E}[Y|X])(\mathbb{E}[Y|X]-Y)|X]]$$

(Keeping just what is needed) $= \mathbb{E}_X[(f(X)-\mathbb{E}[Y|X])\mathbb{E}_{Y|X}[(\mathbb{E}[Y|X]-Y)|X]] = 0$

where the 2nd last step follows since conditioning on X, f(X) and E[Y | X] are constant

- Therefore, $\mathbb{E}[(f(X)-Y)^2] = \mathbb{E}[(f(X)-\mathbb{E}[Y|X])^2] + \mathbb{E}[(\mathbb{E}[Y|X]-Y)^2]$

$$\geq \mathbb{E}[(\mathbb{E}[Y|X]-Y)^2]$$

since the first term being square of a quantity is non-negative

3: Simple Linear Regression

- Consider real-valued variables X and Y. The Y variable is generated, conditional on X, from the following process: $\epsilon \sim N(0, \sigma^2)$ and $Y = aX + \epsilon$ where every ϵ is an independent variable, called a noise term, which is drawn from a Gaussian distribution with mean 0, and standard deviation σ
- The conditional probability of Y has distribution $p(Y \mid X, a) \sim N(aX, \sigma^2)$, so it can be written as

$$p(Y|X,a) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y - aX)^2\right)$$

- Assume we have a training dataset of n pairs (X_i, Y_i) for i = 1,…,n, and σ is known

(a) Frame the maximum likelihood problem for estimating a

(b) Derive the maximum likelihood estimate of the parameter a in terms of the training example X_i's and Y_i's

- Hint: start with the simplest form of the problem you found in the previous question

- Excellent book chapter by Tufte (1974) on one-feature linear regression: `http://www.edwardtufte.com/tufte/dapp/chapter3.html`

3: Solution

(a) $\hat{a}_{MLE}$ = argmax_a P (X, Y | a) =

$$\arg\max_a \prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2)$$

$$\arg\max_a \prod_i \exp(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2)$$

$\hat{a}_{MLELog}$ = $\arg\min_a \frac{1}{2} \sum_i (Y_i - aX_i)^2$

3: Solution

(b) Use $F(a) = \frac{1}{2}\sum_i (Y_i - aX_i)^2$ and minimize F

- Then, ∂ F(α)/ ∂ α = $\sum$ ∂ f(α)/ ∂ α where f(α) = ½ ($Y_i - \alpha X_i$) ^{2}
- If we apply the chain rule $u_1(\alpha) = (Y_i - \alpha X_i)$ and $u_2(\alpha, u_1) = u_1^2$ we obtain $\partial f(\alpha)/\partial \alpha = \partial u_2(\alpha, u_1) / \partial \alpha = \partial u_2(\alpha, u_1) / \partial u1(\alpha) * \partial u1(\alpha) / \partial \alpha = ((Y_i - \alpha X_i)(-X_i)$
- Hence,

$$0 = \frac{\partial}{\partial a}\left[\frac{1}{2}\sum_i (Y_i - aX_i)^2\right] = \sum_i (Y_i - aX_i)(-X_i) = \sum_i aX_i^2 - X_iY_i$$

and

$$\hat{a} = \frac{\sum_i X_iY_i}{\sum_i X_i^2}$$

3 Bonus Exercise: MAP vs MLE Estimation (**40 pts**)

- Let's put a prior on α. Assume $\alpha \sim N(0, \lambda^2)$, so

$$p(a|\lambda) = \frac{1}{\sqrt{2\pi}\lambda} \exp(-\frac{1}{2\lambda^2} a^2)$$

- The posterior probability of a is

$$p(a \mid Y_1, \ldots Y_n, X_1, \ldots X_n, \lambda) = \frac{p(Y_1, \ldots Y_n | X_1, \ldots X_n, a) p(a|\lambda)}{\int_{a'} p(Y_1, \ldots Y_n | X_1, \ldots X_n, a') p(a'|\lambda) da'}$$

- We can ignore the denominator when doing Maximum a Posteriori (MAP) estimation

(c) (**9 points**) Under the following conditions, how do the prior and conditional likelihood curves change?

- Do α_{MAP} and α_{MLE} become closer together, or further apart?

3 Bonus Exercise: MAP vs MLE estimation

	$p(a\|\lambda)$ prior probability: wider, narrower, or same?	$p(Y_1 \dots Y_n \| X_1 \dots X_n, a)$ conditional likelihood: wider, narrower, or same?	$\|a^{MLE} - a^{MAP}\|$ increase or decrease?
As $\lambda \to \infty$			
As $\lambda \to 0$			
More data: as $n \to \infty$ (fixed λ)			

3 Bonus Exercise: MAP vs MLE estimation

(d) (**31 points**) Assume $\sigma = 1$, and a fixed prior parameter λ

- Solve for the MAP estimate of a,

$$\arg\max_{a} \left[\ln p(Y_1..Y_n \mid X_1..X_n, a) + \ln p(a|\lambda) \right]$$

- Your solution should be in terms of X_i's, Y_i's, and λ