MENTOR GUIDE TO LLSE 1

COMPUTER SCIENCE MENTORS 70

November 14 to November 18, 2016

1 Covariance

1.1 Formulas

Provided that X and Y are random variables, we define **covariance** as follows:

$$\text{Cov}(X, Y) = \text{E}((X - \text{E}(X)) \cdot (Y - \text{E}(Y)))$$

Claim: $\text{Cov}(X, Y) = \text{E}(XY) - \text{E}(X)\text{E}(Y)$

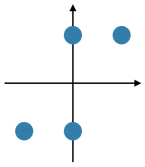
$$\begin{aligned}\text{E}((X - \text{E}(X)) \cdot (Y - \text{E}(Y))) & \\ &= \text{E}(XY - \text{E}(X)Y - X\text{E}(Y) + \text{E}(X)\text{E}(Y)) \\ &= \text{E}(XY) - \text{E}(X)\text{E}(Y) - \text{E}(X)\text{E}(Y) + \text{E}(X)\text{E}(Y)\end{aligned}$$

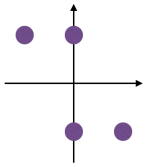
We can cancel out the last two terms:

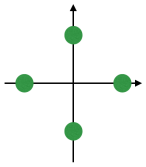
$$= \text{E}(XY) - \text{E}(X)\text{E}(Y)$$

1.2 Intuition

Look at the following graphs. What is the covariance of each? If X increases, what happens to Y? Are these events independent? Does uncorrelated imply independence? Does independence imply uncorrelated?







1.3 Properties of Covariance

1. $\text{Var}(X) = \text{Cov}(X, X)$
2. X, Y independent $\rightarrow \text{Cov}(X, Y) = 0$. Independent random variables are uncorrelated. The converse is not necessarily true (look at the last example on the previous page).
3. $\text{Cov}(a + X, b + Y) = \text{Cov}(X, Y)$
4. $\text{Cov}(aX+bY, cU+dV) = ac\cdot\text{Cov}(X, U)+ad\cdot\text{Cov}(X, V)+bc\cdot\text{Cov}(Y, U)+bd\cdot\text{Cov}(Y, V)$

Proofs

1. Plug into original equation.
2. Plug into original equation (see what you can cancel out by applying the definition of covariance and independence of the variables).
3. Plug into original equation.
4. This one is a little tricky. Since we know that we can add and subtract **constants**, we can zero mean the variables. So we can assume that $\text{E}(aX + bY) = \text{E}(cU + dV) = 0$. Now let's apply the definition of covariance.

   $\text{Cov}(aX + bY, cU + dV)$

   Since $\text{E}(aX + bY) = \text{E}(cU + dV) = 0$, we get

   $$= \text{E}((aX + bY) \cdot (cU + dV)) - 0 \cdot 0$$

   By linearity of expectation:

   $$= ac \cdot \text{E}(XU) + ad \cdot \text{E}(XV) + bc \cdot \text{E}(YU) + bd \cdot \text{E}(YV)$$

   Now note that $\text{Cov}(X, U) = \text{E}(XU) - \text{E}(X)\text{E}(U) = \text{E}(XU) - 0 \cdot 0 = \text{E}(XU)$

   Apply the above result to each term to get:

   $$= ac \cdot \text{Cov}(XU) + ad \cdot \text{Cov}(XV) + bc \cdot \text{Cov}(YU) + bd \cdot \text{Cov}(YV)$$

2 LLSE

2.1 Constant

Motivation: Say we have a random variable Y, which measures height. What would be our best guess for Y? In other words, if a random person called us, what would be our best guess for their height?

2.2 Minimizing Error

We need to first somehow quantify "the best guess". When we choose a number for height, how do we know that the number we chose is a good gues? It makes sense to pick such a number, where the error is minimized. We will look at the **mean square error**,

$$\mathrm{E}((Y-a)^2)$$

and show that this equation achieves a minimum value only when $a = \mathrm{E}(Y)$.

Define

$$\hat{Y} = Y - \mathrm{E}(Y)$$

$\hat{Y}$ is the error we make by using $\mathrm{E}(Y)$ to estimate Y.

Note that,

$$\mathrm{E}(\hat{Y}) = \mathrm{E}(Y - \mathrm{E}(Y)) = \mathrm{E}(Y) - \mathrm{E}(\mathrm{E}(Y)) = \mathrm{E}(Y) - \mathrm{E}(Y) = 0$$

This implies that $\mathrm{E}(\hat{Y} \cdot c) = 0 \forall c$

Now we will use these facts to show that if a is any constant other than $\mathrm{E}(Y)$, the mean square error will be larger than if we had used $\mathrm{E}(Y)$.

$$\begin{aligned}
\mathrm{E}((Y-a^2)) &= \mathrm{E}((Y - \mathrm{E}(Y) + \mathrm{E}(Y) - a)^2) && \text{trick} \\
&= \mathrm{E}((\hat{Y} + c)^2) && \text{where } c = \mathrm{E}(Y) - a \\
&= \mathrm{E}(\hat{Y}^2 + 2\hat{Y}c + c^2) && \text{expand the square} \\
&= \mathrm{E}(\hat{Y})^2 + 2\mathrm{E}(\hat{Y}c) + c^2 && \text{linearity of expectation} \\
&= \mathrm{E}(\hat{Y})^2 + 0 + c^2 && \text{using fact from above} \\
&\geq \mathrm{E}(\hat{Y}^2) && c^2 \geq 0
\end{aligned}$$

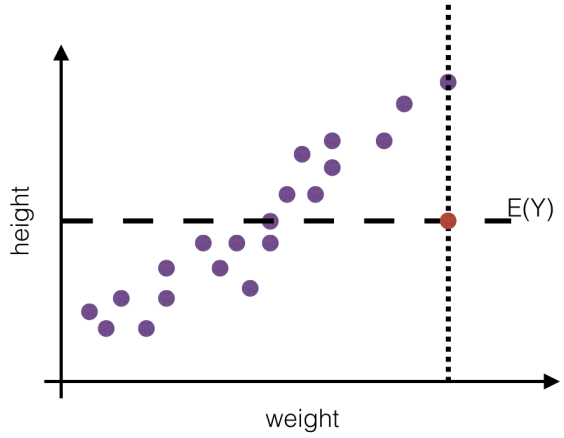
Hence,

$$\mathrm{E}((Y-a)^2) \geq \mathrm{E}((Y - \mathrm{E}(Y))^2)$$

2.3 Linear

We now have a good estimate for Y if we have no additional information. However, what if we know some samples of Y and some other random variable X. For example, say we know the weight of the caller. We should be better able to guess their height, as there is some sort of correlation between weight and height. However, the model we are using so far does not apply any additional information to the guess. Let us construct a new way to predict Y, based on some observed X.

Continuing the example we have been using so far with height and weight, assume we are given the the points shown below.



We can see that $\mathrm{E}(Y)$ is not a good guess given additional information. We wish to construct a linear equation which given a value of X, will predict the value of Y. So if someone calls you and provides their weight, you can more accurately predict what their height is.

So far, our example has been non Bayesian, meaning we did not know the distributions of X and Y. We will develop the linear least square estimate using X, Y with known distributions. Then we will see how we can apply these results to non Bayesian stuff.

Theorem Assume X, Y are random variables with known distribution. Then,

$$\mathrm{L}[Y|X] = \mathrm{E}(Y) + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \cdot (X - \mathrm{E}(X)) = \hat{Y}$$

We want to show that the mean square between Y and $\hat{Y}$ is smaller than between Y and any other linear function of X. We will use a technique similar to how we showed the constant case.

2.4 First Proof of $L[Y|X]$

Recall,

$$\hat{Y} = \mathrm{E}(Y) + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \cdot (X - \mathrm{E}(X))$$

Plug the above equation into the mean value of the error:

$$\begin{aligned} Y - \hat{Y} &= Y - \mathrm{E}(Y) - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \cdot (X - \mathrm{E}(X)) \\ \mathrm{E}(Y - \hat{Y}) &= \mathrm{E}\left(Y - \mathrm{E}(Y) - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \cdot (X - \mathrm{E}(X))\right) \\ &= \mathrm{E}(Y - \mathrm{E}(Y)) - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}\mathrm{E}(X - \mathrm{E}(X)) \\ &= 0 - 0 = 0 \end{aligned}$$

So we know that the mean value of the error is 0.

Claim: $\mathrm{E}((Y - \hat{Y})X) = 0$

Proof of claim:

Note,

$$\mathrm{E}((Y - \hat{Y})X) = \mathrm{E}((Y - \hat{Y}) \cdot (X - \mathrm{E}(X)))$$

To see that the two sides are equivalent we do the following: expand the LHS and apply linearity of expectation we get,

$$\begin{aligned} \mathrm{E}((Y - \hat{Y}) \cdot (X - \mathrm{E}(X))) &= \mathrm{E}(YX - \hat{Y}X - Y\mathrm{E}(X) + \hat{Y}\mathrm{E}(X)) \\ &= \mathrm{E}(YX - \hat{Y}X) - \mathrm{E}(X)\mathrm{E}(Y - \hat{Y}) \qquad \text{mean value of error is 0} \\ &= \mathrm{E}(YX - \hat{Y}X) = \mathrm{E}((Y - \hat{Y})X) \end{aligned}$$

Now simplify the new equation, $\mathrm{E}((Y - \hat{Y}) \cdot (X - \mathrm{E}(X)))$:

$$\begin{aligned} \mathrm{E}((Y - \hat{Y}) \cdot (X - \mathrm{E}(X))) &= \mathrm{E}\left(\left(Y - \left(\mathrm{E}(Y) + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \cdot (X - \mathrm{E}(X))\right) \cdot (X - \mathrm{E}(X))\right) \\ &= \mathrm{E}((Y - \mathrm{E}(Y))(X - \mathrm{E}(X))) - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}\mathrm{E}((X - \mathrm{E}(X))(X - \mathrm{E}(X))) \\ &= \mathrm{Cov}(X,Y) - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}\mathrm{Var}(X) \\ &= 0 \end{aligned}$$

We conclude that $\mathrm{E}((Y - \hat{Y})X) = 0$.

We need one more fact in order to prove our original claim:

Claim: The error is orthogonal to all linear functions. (We have shown that it is orthogonal to x)

Proof of claim:

$$\begin{aligned}
\mathrm{E}((Y-\hat{Y})(c+dX)) &= \mathrm{E}(Yc - \hat{Y}c + YdX - \hat{Y}dX) \\
&= c\mathrm{E}(Y-\hat{Y}) + d\mathrm{E}((Y-\hat{Y})X) \\
&= 0 - 0 = 0 && \text{using above results}
\end{aligned}$$

Now let's get back to our original claim: We want to show that the mean square between $\hat{Y}$ and Y is smaller than between Y and any other linear function. Recall that to show $\mathrm{E}(Y)$ is the best constant estimate for Y we proved the following equation:

$$\mathrm{E}((Y-a)^2) \geq \mathrm{E}((Y-\mathrm{E}(Y))^2)$$

To show that $\hat{Y}$ is the best linear estimator for Y, we will prove that

$$\mathrm{E}((Y-a-bX)^2) \geq \mathrm{E}((Y-\hat{Y})^2)$$

Note the following relation:

$$\mathrm{E}((Y-\hat{Y})(\hat{Y}-a-bX)) = 0 \forall a, b$$

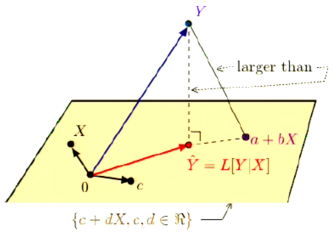
We have shown that error is orthogonal to all linear functions. $\hat{Y}$ and $-a-bX$ are both linear functions, so adding them produces a linear function.

Now we use this fact to prove the inequality:

$$\begin{aligned}
\mathrm{E}((Y-a-bX)^2) &= \mathrm{E}((Y-\hat{Y}+\hat{Y}-a-bX)^2) \\
&= \mathrm{E}((Y-\hat{Y})^2) + \mathrm{E}((\hat{Y}-a-bX)^2) + 0 \\
&\geq \mathrm{E}((Y-\hat{Y})^2) && \text{as } \mathrm{E}((\hat{Y}-a-bX)^2) \geq 0
\end{aligned}$$

We have shown that the mean square error from using $a+bX$ is always going to be larger than if we used $\hat{Y}$.

The diagram below offers a visualization of the above facts (from Walrand)



It also demonstrates the **Projection Property**. If the plane is the set of all linear functions of X, then $Y - \hat{Y}$ is orthogonal (perpendicular) to any linear function of X (or the plane).

Here is another derivation of $\hat{Y}$ using a calculus approach.

2.5 Second Proof of $L[Y|X]$

Assume that $\mathrm{E}(X) = \mathrm{E}(Y) = 0$.

$$\hat{Y} = \mathrm{E}(Y) + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \cdot (X - \mathrm{E}(X)) = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \cdot X$$

We want to find a and b such that $g(a,b)$ is minimized, where $\hat{Y} = a + bX$

$$\begin{aligned} g(a,b) &= \mathrm{E}((Y - a - bX)^2) \\ &= \mathrm{E}(Y^2 + a^2 + b^2X^2 - 2aX - 2bXY + 2bX) \\ &= a^2 + \mathrm{E}(Y^2) + b^2\mathrm{E}(X^2) - 2a\mathrm{E}(Y) - 2b\mathrm{E}(XY) + 2b\mathrm{E}(X) \\ &= a^2 + \mathrm{E}(Y^2) + b^2\mathrm{E}(X^2) - 2b\mathrm{E}(XY) \qquad \text{X, Y are zero mean} \end{aligned}$$

To find the minimum value of $g(a,b)$ take the partial derivatives with respect to a and b and set them both to zero. Then solve to find a and b.

$$0 = \frac{\partial}{\partial a} g(a,b) = 2a \rightarrow a = 0$$

$$0 = \frac{\partial}{\partial b} g(a,b) = 2b\mathrm{E}(X^2) - 2\mathrm{E}(XY) \rightarrow b = \frac{\mathrm{E}(XY)}{\mathrm{E}(X^2)} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$$

The last line follows from the fact that X and Y are zero mean:

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = \mathrm{E}(XY) - 0 \cdot 0 = \mathrm{E}(XY)$$
$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = \mathrm{E}(X^2) - 0 = \mathrm{E}(X^2)$$

$$\hat{Y} = a + bX = 0 + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \cdot X$$

This proves that $\hat{Y}$ is the best linear estimate for Y.

In the above proof we forced X and Y to be zero mean. We can use a similar proof technique with any X and Y.

2.6 Third Proof of $L[Y|X]$

This applies the second proof to X and Y that are not zero mean. This proof is not covered in lecture and basically involves maneuvering terms around, so read it if you're curious.

$$\begin{aligned} Y - a - bX &= Y - \mathrm{E}(Y) - (a - \mathrm{E}(Y)) - b(X - \mathrm{E}(X)) - b\mathrm{E}(X) && \text{add some magic terms} \\ &= Y - \mathrm{E}(Y) - (a - \mathrm{E}(Y) + b\mathrm{E}(X)) - b(X - \mathrm{E}(X)) && \text{rearrange terms} \\ &= Y - \mathrm{E}(Y) - c - b(X - \mathrm{E}(X)) && c = a - \mathrm{E}(Y) + b\mathrm{E}(X) \end{aligned}$$

Now apply results from the Second Proof of $L[Y|X]$, since $Y - \mathrm{E}(Y)$ and $X - \mathrm{E}(X)$ are both zero mean.

$$c = 0 \to 0 = a - \mathrm{E}(Y) + b\mathrm{E}(X) \to a = \mathrm{E}(Y) - b\mathrm{E}(X)$$
$$b = \frac{\mathrm{Cov}(X - \mathrm{E}(X), Y - \mathrm{E}(Y))}{\mathrm{Var}(X - \mathrm{E}(X))} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}$$

Now plug in the above results into $a + bX$:

$$\begin{aligned} a + bX &= \mathrm{E}(Y) - b\mathrm{E}(X) + bX \\ &= \mathrm{E}(Y) - b(-\mathrm{E}(X) + x) = \mathrm{E}(Y) + \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}(X - \mathrm{E}(X)) \end{aligned}$$

2.7 Some General Comments about LLSE

$$\mathrm{L}[Y|X] = \mathrm{E}(Y) + \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} \cdot (X - \mathrm{E}(X))$$

1. Covariance controls the slope of LLSE (variance is always positive). If the covariance is positive, then as X increases, our prediction for Y will also increase.
2. If you know the distributions of X and Y, you can calculate each component of the above equation.
3. If X and Y are independent, then the best guess is $\mathrm{E}(Y)$
4. If we have no observations, the best choice for an estimate of Y is $\mathrm{E}(Y)$. In this case, the mean error would be:
$$\mathrm{E}((Y - \mathrm{E}(Y))^2) = \mathrm{Var}(Y)$$
5. Observing X will reduce the error. The error depends on how X and Y are correlated and is defined as follows:
$$\mathrm{E}((Y - \mathrm{L}[Y|X] = \mathrm{Var}(Y) - \frac{\mathrm{Cov}(X, Y)^2}{\mathrm{Var}(X)}$$
Note that both the numerator and denominator in the fraction are nonnegative. Since the error of using $\mathrm{E}(Y)$ is $\mathrm{Var}(Y)$, the error of using $\mathrm{L}[Y|X]$ will always be smaller than if we just used $\mathrm{E}(Y)$.

2.8 Bayesian and non Bayesian

We motivated this section with a non Bayesian example. Then we used Bayesian stuff in all of our proofs. Now we explain that Bayesian and non Bayesian are really the same thing.

non Bayesian

Given samples: $\{(X_n, Y_n)\, n = 1, \ldots N\}$ We want

$$\hat{Y} = a + bX$$

where a, b minimize

$$\sum_{n=1}^{N} (Y_n - a - bX_n)^2$$

In other words, we minimize the square of the distance between the line and the points. We take the square of the distance because taking the absolute value of the difference is more complicated.

Note that in the above formulation, we made no assumptions about the distribution of X.

Bayesian

Given X, Y with known distribution. We want:

$$\hat{Y} = a + bX = \mathrm{L}[Y|X]$$

where

$$g(a, b) = \mathrm{E}((Y - a - bX)^2)$$

is minimized.

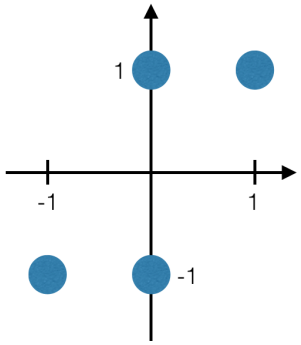
We can turn non Bayesian into Bayesian as follows:

$$\frac{1}{N} \sum (Y_n - a - bX_n)^2 = \mathrm{E}((Y - a - bX)^2)$$

$$(X, Y) = (X_n, Y_n) \text{ w. p. } \frac{1}{N} \text{ for } n = 1, \ldots, N$$

3 Example of Finding LLSE (non Bayesian)

Find the linear least square estimate of the following joint distribution.



We turn the problem from non Bayesian into Bayesian by assuming that the 4 points are equally likely values of (X, Y).

$$\begin{aligned}
\mathrm{E}(X) &= -1 \cdot \frac{1}{4} + 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} \\
\mathrm{E}(Y) &= -1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = 0 \\
\mathrm{E}(X^2) &= (-1)^2 \cdot \frac{1}{4} + (0)^2 \cdot \frac{1}{2} + (1)^2 \cdot \frac{1}{4} = \frac{1}{2} \\
\mathrm{E}(XY) &= 1 \cdot \frac{1}{2} = \frac{1}{2} \\
\mathrm{Var}(X) &= \mathrm{E}(X^2) - \mathrm{E}(X)^2 = \frac{1}{2} - 0 = \frac{1}{2} \\
\mathrm{Cov}(X, Y) &= \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = \frac{1}{2} - 0 = \frac{1}{2}
\end{aligned}$$

Therefore we get:

$$\mathrm{L}[Y|X] = 0 + \frac{\frac{1}{2}}{\frac{1}{2}} \cdot X = X$$

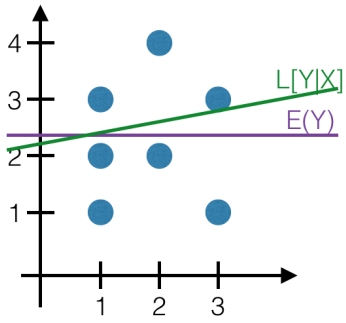
4 Conditional Expectation

Motivation

There are situations where the relationship between X and Y is non linear. Our goal is to find $g(\cdot)$ such that $g(X)$ is the best guess about Y given X where $g(\cdot)$ is not restricted to a line.

4.1 Motivating Example

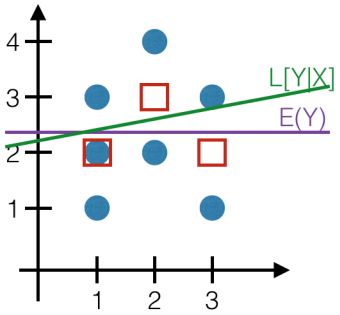
Examine the following joint distribution of X and Y.



Without any observations, our best guess for Y is $\mathrm{E}(Y) = 2.3$. Assuming we observe X, we can construct the following linear estimate:

$$\mathrm{L}[Y|X] = a + bX = 2.1 + 0.1X$$

What if we know that X is 1? What would be our best guess for Y in that case? 2! Since there are 3 values that Y can take on one we know that X is 1, we find the average of those values. So once we narrow down X, we examine a new distribution. Repeating this process for each value of X gives us the following diagram:



Compare the each of the new expected values with what the linear equation gives us. Which is closer?

4.2 Defining Conditional Expectation

If X and Y are random variables in Ω then the **conditional expectation** is defined as follows:

$$\begin{aligned} \mathrm{E}(Y|X) &= g(X) \\ \text{where, } g(X) &= \mathrm{E}(Y|X=x) \\ &= \sum_y y \cdot \mathrm{P}[Y=y|X=x] \\ &\text{with } \mathrm{P}[Y=y|X=x] = \frac{\mathrm{P}[X=x, Y=y]}{\mathrm{P}[X=x} \end{aligned}$$

Theorem

$\mathrm{E}(Y|X)$ is the best guess about Y given X.

$$\mathrm{E}((Y-g(X))^2) \geq \mathrm{E}((Y-\mathrm{E}(Y|X))^2)$$

Using any other function to predict Y will give a larger error.

4.3 Projection Property Revisited

Projection Property: $\mathrm{E}((Y-\mathrm{E}(Y|X))f(X)) = 0 \forall f(\cdot)$

This is equivalent to the following,

$$\begin{aligned} \mathrm{E}((Y-\mathrm{E}(Y|X))f(X)) &= \mathrm{E}(Y \cdot f(X)) - \mathrm{E}(\mathrm{E}(Y|X) \cdot f(X)) \\ &\rightarrow \\ \mathrm{E}(Y \cdot f(X)) &= \mathrm{E}(\mathrm{E}(Y|X) f(X)) \end{aligned}$$

Proof:

$$\begin{aligned} \mathrm{E}(\mathrm{E}(Y|X) \cdot f(X)) &= \sum_x \mathrm{E}(Y|X=x) \cdot f(x) \cdot \mathrm{P}[X=x] \\ &= \sum_x \left(\sum_y y \cdot f(x) \cdot \mathrm{P}[Y=y|X=x] \right) \cdot \mathrm{P}[X=x] \\ &= \sum_x \sum_y y \cdot f(x) \mathrm{P}[X=x, Y=x] \qquad \text{using } \frac{\mathrm{P}[X=x, Y=y]}{\mathrm{P}[X=x]} \\ &= \mathrm{E}(Y \cdot f(X)) \end{aligned}$$

Therefore, if we have the conditional expectation, we can take its mean value and get the mean value of Y.

4.4 CE = MMSE

Finally we prove the fact that $\mathrm{E}(Y|X)$ is the best guess about Y based on X. We use a technique similar to before.

Recall,

$$\begin{aligned}\mathrm{E}(Y|X) &= g(X)\\ \text{where, } g(X) &= \mathrm{E}(Y|X=x)\\ &= \sum_y y \cdot \mathrm{P}[Y=y|X=x]\\ &\text{with } \mathrm{P}[Y=y|X=x] = \frac{\mathrm{P}[X=x, Y=y]}{\mathrm{P}[X=x}\end{aligned}$$

We want to show that $g(X)$ minimizes $\mathrm{E}((Y-g(X))^2)$. Let $h(X)$ be any other function.

$$\begin{aligned}\mathrm{E}((Y-h(X))^2) &= \mathrm{E}((Y-g(X)+g(X)-h(X))^2) && \text{add in some magic terms}\\ &= \mathrm{E}((Y-g(X))^2) + \mathrm{E}((g(X)-h(X))^2)\\ &+ 2\mathrm{E}((Y-g(X))(g(X)-h(X)) && \text{linearity of expectation}\end{aligned}$$

We know that $\mathrm{E}((Y-g(X))(g(X)-h(X))=0$ by the Projection Property and $\mathrm{E}((g(X)-h(X))^2 \geq 0$.

Therefore,

$$\mathrm{E}((Y-h(X))^2) \geq \mathrm{E}((Y-g(X))^2)$$