Notes on Probabilistic Estimations

Davis Yang Email:yxyang@berkeley.edu

November 15, 2016

The past week of CS70 talked a lot about estimation using probability. Considering that the concepts including LLSE, MMSE, Regression and stuff can be quite confusing, I'm writing this note to summarize various concepts and explain how they are connected, in the hope of giving a general picture.

The note starts with MMSE, which is the most general form of estimation. Then it moves to talk about LLSE and Quadratic Estimation. In the end, it talks about regression problems and shifts from a conceptual point of view to a more practical understanding.

1 Motivating Example

Let's start with the following example. [1]

Example. You have a fair, 6-sided dice and a fair coin. You first rolled the dice, and say the dice comes up with number x. Then you flip the coin 2^x number of times, and count the number of heads that comes up.

Let's denote random variable X as the outcome of dice roll, and Y as the number of heads that show up. For example, you rolled the dice and it comes up with number 4, then you proceed to flip the coin $2^4 = 16$ times, and get 7 heads. Then, for this experiment, we have $X = 4, Y = 7$.

Now, suppose your friend knows X, the result of dice roll. And he wants to guess the number of heads you get. Which number should he guess?

This is a good example of *probabilistic estimation* [2]. You're given a bunch of random variables and values of a subset of them, and you want to estimate the value of the rest. Plus, you know the joint distribution of all variables (this is a strong assumption, and we'll relax it in the last portion of this note). But for now, let's stick to the previous example: you know the joint distribution of X and Y, and the value of X, you want to give the best estimate of Y.

2 MMSE

2.1 An evaluation metric first

We've specified the problem, but haven't specified what a good "estimation" is yet. For this note[3], let's stick to the following error metric:

[1] Forgive me if it sounds boring, I could come up with more interesting and real-life ones, but explaining them would be a lot harder.

[2] this is technically not a term, but I use it to differentiate from regression

[3] In fact, there are other error terms being used, beyond the scope of this course

Definition. Suppose your estimation of Y's value given observation of X is $g(X)$. Then the **mean square error** of Y and $g(X)$ is defined as:

$$E[(Y - g(X))^2]$$

Let's explore this metric more carefully:

$$E[(Y - g(X))^2] = \sum_x \sum_y (y - g(x))^2 P(X = x, Y = y)$$

For all possible combinations of X and Y, we calculate the squared difference between the actual value of Y and estimated value $g(X)$. The use of square prevents positive and negative errors from being canceled. It also makes it easier to take the derivative. Then we take the expectation of it, which can be interpreted as a "weighted average".

2.2 MMSE — A general estimation scheme

MMSE, as the name "minimum mean squared estimate" implies, minimizes the error term described in section 2.1. While deriving the MMSE function is nontrivial, it's pretty easy to show that having $g(X) = E[Y|X]$ minimizes the mean squared error [4]

Theorem. Let X and Y be random variables, then the MMSE of Y given X is $E[Y|X]$. i.e. for any other function $g(x)$,

$$E[(Y - E[Y|X])^2] \leq E[(Y - g(X))^2]$$

Proof. Let's use linearity of expectation to expand the mean squared error expression:

$$\begin{aligned} E[(Y - g(X))^2] &= E\Big[\Big(Y - E[Y|X] + E[Y|X] - g(X)\Big)^2\Big] \\ &= E\Big[(Y - E[Y|X])^2 + 2(Y - E[Y|X])(E[Y|X] - g(X)) + (E[Y|X] - g(X))^2\Big] \\ &= E[(Y - E[Y|X])^2] + 2E[(Y - E[Y|X])(E[Y|X] - g(X))] \\ &+ E\Big[\big(E[Y|X] - g(X)\big)^2\Big] \end{aligned}$$

Note that the first term has no dependence on g and can be treated as constant. The second term is 0 following the projection property of conditional expectation. [5]. The third term is nonnegative and is zero iff $g(X) = E[Y|X]$ Therefore, choosing $g(X) = E[Y|X]$ minimizes the mean squared error. □

Back to the example defined in section 1, to calculate the MMSE given Y, we need to calculate $E[Y|X]$. Note that given $X = x$, the fair coin is flipped 2^x times, and we're counting the number of heads showing up, which is just the expectation of a binomial random variable. Therefore, $E[Y|X = x] = \frac{1}{2} \cdot 2^x = 2^{x-1}$.

[4] Remember, $E[Y|X] = \sum_y yP(Y = y|X = x)$ is the expected value of Y given X, which is a function of X

[5] Refer to lecture 31 slide page 10 for projection property

3 LLSE and Generalized LLSE

What we do in LLSE (which stands for linear least square estimate), is really not too different from MMSE. We again look for a function $g(X)$ to estimate Y. We also use the same mean squared error as the evaluation metric. However there's one thing in addition: we require $g(X)$ to be linear in X: i.e. $g(X) = aX + b$

3.1 Basic, 2-variable LLSE

We start by presenting the formula for LLSE when there's only one independent variable X and one dependent variable Y.

Theorem. Let X and Y be random variables, the LLSE of Y given X,

$$L(Y|X) = E[Y] + \frac{cov(X,Y)}{var(X)}(X - E[X])$$

minimizes the mean squared error. That is, for any other *linear* function $g(X) = aX + b$,

$$E[(Y - L(Y|X))^2] \leq E[(Y - g(X))^2] = E[(Y - aX - b)^2]$$

Proof. Let's first show 2 important result. $E\Big[\big(Y - L(Y|X)\big)\Big] = 0$ and $E\Big[X\big(Y - L(Y|X)\big)\Big] = 0$

$$\begin{aligned}
E[L(Y|X)] &= E[E[Y] + \frac{cov(X,Y)}{var(X)}(X - E[X])] \\
&= E[Y] + \frac{cov(X,Y)}{var(X)}E\big[X - E[X]\big] \\
&= E[Y] + \frac{cov(X,Y)}{var(X)}(E[X] - E[X]) \\
&= E[Y] \implies E[Y - L(Y|X)] = 0 \\
E\big[XL(Y|X)\big] &= E\Big[X\big(E[Y] + \frac{cov(X,Y)}{var(X)}(X - E[X])\big)\Big] \\
&= E\left[XE[Y] + \frac{cov(X,Y)}{var(X)}X^2 - \frac{cov(X,Y)}{var(X)}XE[X]\right] \\
&= E[X]E[Y] + \frac{cov(X,Y)}{var(X)}(E[X^2] - E[X]^2) \\
&= E[X]E[Y] + cov(X,Y) \\
\implies E\Big[X\big(Y - L(Y|X)\big)\Big] &= E[XY] - (E[X]E[Y] + cov(X,Y)) = 0
\end{aligned}$$

Now, let's apply the trick in MMSE again. For any other linear function $g(x) = ax + b$, we have:

$$\begin{aligned}
E[(Y - g(X))^2] &= E[(Y - aX - b)^2] \\
&= E\Big[\big((Y - L(Y|X)) + (L(Y|X) - aX - b)\big)^2\Big] \\
&= E\Big[\big(Y - L(Y|X)\big)^2\Big] + 2E\Big[\big(Y - L(Y|X)\big)\big(L(Y|X) - aX - b\big)\Big] \\
&\quad + E[(L(Y|X) - aX - b)^2]
\end{aligned} \tag{1}$$

Note that since $L(Y|X)$ is also linear in X. Therefore, we have

$$L(Y|X) - aX - b = cX + d$$

where c, d are also constants. Therefore,

$$\begin{aligned} E\Big[\big(Y - L(Y|X)\big)\big(L(Y|X) - aX - b\big)\Big] &= E[(Y - L(Y|X))(cX + d)] \\ &= cE[(Y - L(Y|X))X] + dE[Y - L(Y|X)] \\ &= c \cdot 0 + d \cdot 0 = 0 \end{aligned}$$

Therefore, looking back at equation (1), in the final summation, the first term is unrelated to a, b, the second term is 0 and the third term is nonnegative and is zero iff $g(X) = aX+b = L(Y|X)$ $\square$

Back to the example of dice roll + coin flip, we calculate (via NumPy) that

$$cov(X, Y) = 16.75, var(X) = \frac{35}{12}, E[X] = 3.5, E[Y] = 10.5$$

Therefore, we have: $L(Y|X) = 10.5 + \frac{201}{35}(X - 3.5)$

3.2 Extension: LLSE with more variables

Okay we just found the LLSE of Y given X. What if the value of Y is related not only to X, but also some other random variables? For example, if the coin is no longer fair and instead have head probability P uniformly distributed between $0.3, 0.6, 0.9$ and you are trying to estimate Y given the value of X and P?

Intuitively, the MMSE shouldn't change much: it should still $E[Y|X, P]$. What about LLSE in this case? i.e., how do we find coefficient a, b, c such that $g(X, P) = aX + bP + c$ best estimates Y in terms of mean squared error?

It's difficult to find a closed-form solution like the one in section 3.1 in this case, but the general idea is still the same. Specifically, if we find $g(X, P) = aX + bP + c$ such that:

$$\begin{cases} E[X(Y - aX - bP - c)] = 0 \\ E[P(Y - aX - bP - c)] = 0 \\ E[1(Y - aX - bP - c)] = 0 \end{cases}$$

We can prove (feel free to do this as a practice) that the corresponding $g(X, P) = aX + bP + c$ minimizes the mean squared error. You can either follow similar proof idea of the previous section. Another way to see this is by setting the derivative of $E[(Y - aX - bP - C)^2]$ w.r.t a, b, c to 0. For example, taking the derivative w.r.t a gives you the first equation:

$$\begin{aligned} \frac{\partial}{\partial a} E[(Y - aX - bP - c)^2] &= \frac{\partial}{\partial a} \sum_{x,p} (Y - ax - bp - c)^2 P(X = x, P = p) \\ &= \sum_{x,p} P(X = x, P = p) \frac{\partial}{\partial a} (Y - ax - bp - c)^2 \\ &= \sum_{x,p} P(X = x, P = p)(-2x)(Y - ax - bp - c) \\ &= -2E[X(Y - aX - bP - C)] = 0 \end{aligned}$$

3.3 Extension: higher-order terms

You may wonder: is a linear estimator a good estimator for the roll-then-toss example in the beginning? Probably not: as the MMSE turns out to be exponential in X, a linear function can't approximate it very well. What if we introduce some higher-order terms? Say $g(X) = cX^2 + aX + b$?

Solving for the optimal coefficient c, a, b turns out to be not too different. We only need to solve (feel free to verify this yourself):

$$\begin{cases} E[X^2(Y - cX^2 - aX - b)] = 0 \\ E[X(Y - cX^2 - aX - b)] = 0 \\ E[1(Y - cX^2 - aX - b)] = 0 \end{cases}$$

Note that there's nothing specific about X^2 being quadratic in the equations above. We can replace X^2 with any arbitrary function of X and solve for the coefficients in a similar fashion.

Back to the example of roll+flip in chapter 1, the best quadratic equation can be solved as:

$$Y = 1.76785714X^2 + -6.63214286X + 6.9$$

3.4 Comparison between MMSE, LLSE and Generalized LLSE

Let's take a look back at the original example of roll-and-flip. We've seen 3 estimates so far: MMSE, LLSE and LLSE with quadratic term. They're all minimizing the same mean squared error. Let's summarize the result:

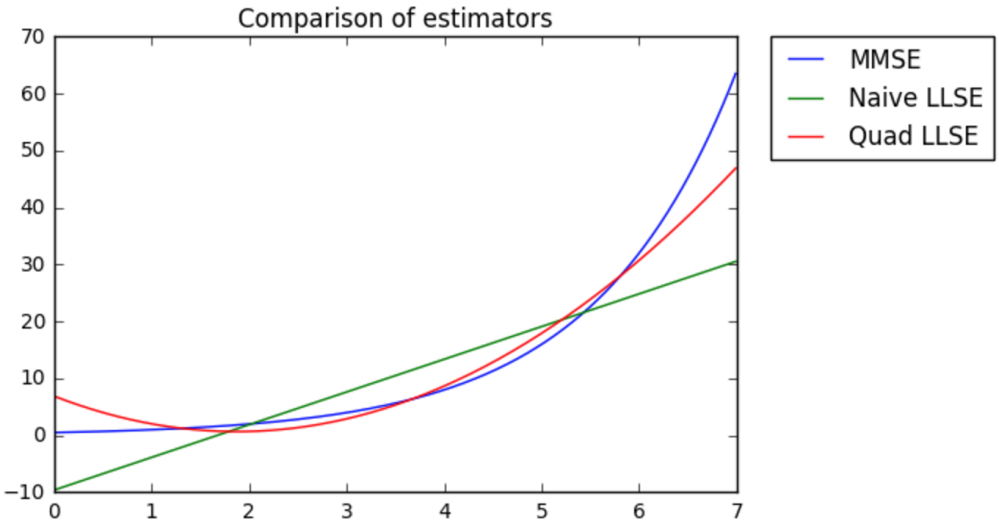
Estimator	Function	Mean Squared Error
MMSE	2^{x-1}	0.2276
LLSE	$10.5 + \frac{201}{35}(X - 3.5)$	3.2557
LLSE w/ Quadratic Term	$Y = 1.76785714X^2 + -6.63214286X + 6.9$	0.4557

The errors should not be too much of a surprise: MMSE is the best possible function that minimizes the mean squared error, as proved in the previous section. Depending on the distribution, the MMSE can be any arbitrary function.

LLSE also tries to minimize the mean squared error, but we're restricted to linear functions. Here since the MMSE is an exponential function, a linear function can't approximate it too well.

Adding a quadratic term improves the error a lot, as we can approximate the exponential function a lot better with the quadratic term. However, we're still unable to meet the performance of MMSE.

Things become more clear when we plot the three functions:



The blue line, which is the optimal MMSE prediction, is way more closely approximated by the quadratic LLSE than the naive version.

3.5 Conclusion

We introduced the basic model of LLSE with 1 independent and 1 dependent variable variable in 3.1. In 3.2, we generalize to the situation where multiple random variables collectively predict Y in a linear fashion. In 3.3 we also discussed how we can introduce any arbitrary function of a variable into the prediction function. 3.2 and 3.3 collectively forms the basic idea of *generalized linear estimation*.

4 Let's talk about regression

Okay, we've seen that MMSE is the best estimator that minimizes the mean squared error. Then why do we need to introduce LLSE, and all those generalizations, which could only perform worse than MMSE? So here's the story:

We've assumed that we know the joint distribution of X and Y throughout the previous chapters. The characterization of roll-and-flip nature in the example completely determines how X and Y are distributed: what value they can take, and the probability of each value. With that, MMSE does work.

Real world is different. In many cases we don't know the real distribution, and the only thing we have is data, which can be thought of as samples of the real distribution. Consider you're predicting the price of a house Y given the number of bedroom it has X. Y and X should probably be positively correlated, but it's really hard to know the joint distribution of X and Y: how the market prices each house considering the number of bedroom it has. In this case MMSE doesn't work.

What can we do? While we don't know the actual distribution of X and Y, we do have records of houses sold and how many bedrooms they had. The records can be thought of as *samples* of the real distribution and should still help us somehow. Assuming that each sample is equally

probable, it's reasonable to assume that if the number of samples is large enough, the sampled distribution should resemble the actual one (think about law of large numbers).

We know that the house price should probably be positively correlated with the number of bedrooms, and the simplest model of positive relationship would be a line with positive slope. So why not try that? Assuming the sampled distribution where each sample is equally likely, we can find a LLSE out of the sampled distribution and see how well it works. We can try to introduce higher-order terms to see if it fits any better... (and the process continues, which is an important part of machine learning, CS188, CS189...)

Note that I've avoided the term **regression** throughout this note. This is because technically, regression assumes no knowledge of the actual distribution and works with samples. Each sample is treated to be equally likely, and we fit some function to all the samples. This is where the "non-Bayesian" or "uniform" nature of linear regression comes from.