CS 70 Discrete Mathematics and Probability Theory
Spring 2025 Course Notes **Note 20**

Prediction!

Prediction is one of the most compelling applications of probability, particularly in computer science.

In our setting, prediction is the problem of making an estimate for a random variable from available information; today that information is a distribution and joint distribution. The goal of prediction is to minimize the error (sometimes called loss) in the prediction in a formal sense.

Today, we consider the *mean squared error*. That is, for a random variable X with a known mean, we wish to give an estimate for x which minimizes the following expression:

$$\mathbb{E}\left[(X-x)^2\right].$$

The solution to this prediction problem is $\mathbb{E}[X]$!

1 The parabola

Mathematically, we revisit those childhood days of understanding the parabola. Recall that for a parabola defined by $f(x) = Ax^2 + Bx + C$, the vertex of the parabola is at $x = -\frac{B}{2A}$. Long ago, you derived this by "completing the square".

Exercise. Try this at home!

One can also use calculus and set the derivative, $f'(x) = 2Ax + B$, to zero, and solving for x to find a critical point. It is a minimum when A is positive (open upwards graphically) and a maximum when A is negative.

We remark that this middle school notion of minimizing a quadratic is very powerful indeed it is a core idea in optimization (see, for example, EECS 127) in general not just in optimal prediction.

2 Expected value is optimal!

Recall, we wish to find the value for x which minimizes $\mathbb{E}\left[(X-x)^2\right]$. Let's do this!

$$\begin{aligned}\mathbb{E}\left[(X-x)^2\right] &= \mathbb{E}\left[X^2\right] - 2\,\mathbb{E}[X]\,\mathbb{E}[x] + \mathbb{E}\left[x^2\right]\\ &= \mathbb{E}\left[X^2\right] - 2\,\mathbb{E}[X]x + x^2\end{aligned}$$

The second line follows from the fact that x is a constant with respect to the expectation. We see that this is a parabola of the form $Ax^2 + Bx + C$ where $A = 1$, $B = -2\,\mathbb{E}[X]$, and $C = \mathbb{E}\left[X^2\right]$. Thus, the minimum is at $x = -\frac{-2\mathbb{E}[X]}{2} = \mathbb{E}[X]$. Middle school is cool.[1]

[1] The joke is that one learns about the parabola in middle school. Unfortunately, middle school is anything but fun. Also, unfortunately, having to explain a joke perhaps means it is a poor one.

An important observation is the minimum value of the mean squared error of this estimator is exactly the variance of X, i.e., $\mathbb{E}\left[(X-\mathbb{E}[X])^2\right]$.

In sum, the best mean squared error of estimator for X is $\mathbb{E}[X]$ and this estimate results in a mean squared error of $\mathrm{Var}(X)$.

3 Joint Distributions: conditional expectation

Recall the following definition that we presented previously.

Definition 20.1. *The joint distribution for two discrete random variables X and Y is the collection of values $\{((a,b),\mathbb{P}[X=a,Y=b]) : a \in \mathscr{A}, b \in \mathscr{B}\}$, where $\mathscr{A}$ is the set of all possible values taken by X and $\mathscr{B}$ is the set of all possible values taken by Y.*

When given a joint distribution for X and Y, the distribution $\mathbb{P}[X=a]$ for X is called the *marginal distribution* for X, and can be found by "summing" over the values of Y. That is,

$$\mathbb{P}[X=a] = \sum_{b\in\mathscr{B}} \mathbb{P}[X=a,Y=b].$$

The marginal distribution for Y is analogous, as is the notion of a joint distribution for any number of random variables.

From a joint distribution, we can compute the *conditional probability* of $X=x$ given that $Y=y$ from the definition of conditional probability as follows:

$$\mathbb{P}[X=x \mid Y=y] = \frac{\mathbb{P}[X=x,Y=y]}{\mathbb{P}[Y=y]}.$$

The *conditional expectation* of X given $Y=y$ is defined naturally as follows:

$$\mathbb{E}[X \mid Y=y] = \sum_{x\in\mathscr{A}} x\cdot\mathbb{P}[X=x \mid Y=y].$$

That is, $\mathbb{E}[X \mid Y=y]$ is simply the expectation of X given that $Y=y$. This is useful for prediction in the same sense that expectation is useful as we will return to later in this note.

3.1 Iterated Expectation and Wald's identity

Before returning to prediction, we discuss *the law of iterated expectations* which is

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]] = \sum_{y\in\mathscr{B}} \mathbb{E}[X \mid Y=y]\,\mathbb{P}[Y=y]. \tag{1}$$

This simple concept can be quite useful. For example, consider choosing an integer N at random and forming a random variable $Y = X_1 + \cdots + X_N$ where the X_i are identical and independently distributed. Note the *number* of terms is a random variable here! We wish to compute $\mathbb{E}[Y]$ and assuming that the random variables X_i is independent of the value of N.

$$\begin{aligned}\mathbb{E}[Y] &= \mathbb{E}[\mathbb{E}[Y \mid N]] \\ &= \sum_n \mathbb{E}[Y \mid N = n]\, \mathbb{P}[N = n] \\ &= \sum_n n\, \mathbb{E}[X_1]\, \mathbb{P}[N = n] \\ &= \mathbb{E}[X_1] \sum_n n\, \mathbb{P}[N = n] \\ &= \mathbb{E}[X_1]\, \mathbb{E}[N]\end{aligned}$$

The third line follows from $Y = X_1 + \cdots + X_n$ and the fact that X_1 are identically distributed and are independent of the value of N. Thus, we have $\mathbb{E}[Y] = \mathbb{E}[X_1]\,\mathbb{E}[N]$ which is the basic form of *Wald's identity.*

This can be useful for modelling the total time to serve customers in a time interval, where we have "Poisson" arrivals, and each customer's service time is from the same distribution. That is, we have a Poisson random variable, $N \sim \text{Poisson}(\lambda)$, that determines the number of customers and each X_i is a random variable that corresponds to the time needed to serve customer i.

To conclude, we note that the law of iterated expectations is sometimes called the tower rule as one can extend the concept to more than two random variables, e.g., $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[\mathbb{E}[X \mid Y, Z]]]$, where the outer expectations are over the values of Y and Z analogously to (1).

4 Minimum Mean Square Estimate (MMSE): an example

Returning to prediction: one predicts given more information just than one's prior expectations or (in our words) distributions to predict a value. We do this constantly in our behavior, e.g., when will a traffic light turn red after we see it turn yellow?

A simpler example is to predict the weight of someone from their height. Here the information is two fold, the first being the joint distribution of the weights, W, and heights, H. That is, we have $\mathbb{P}[W = w, H = h]$. The second being the value of height itself, h.

Exercise. What estimate should one use for W to minimize mean squared error if one does not have a height?

For the known value of h, we wish to find w that minimizes the mean squared error which is:

$$\mathbb{E}\left[(W - w)^2 \mid H = h\right].$$

Recall, that conditioning on $H = h$, in turn yields a distribution on W which we denote as $\mathbb{P}[W = w \mid H = h]$, which in turn has a conditional expectation, denoted by $\mathbb{E}[W \mid H = h]$. This is just the expectation of W in the sample space corresponding to the event $H = h$, and thus by the previous section is the best estimate for W with regards to minimizing the mean squared error.

Note that $\mathbb{E}[W \mid H = h]$ is a function of h, and it is the expectation on the sample space restricted to those with $H = h$. And since for each h, the mean squared error is minimized by the expectation, we have that the mean squared error $\mathbb{E}_{W,H}[(W - \mathbb{E}[W \mid H])^2]$ is minimized with respect to the entire sample space. [2] That is, the expected squared error over the entire joint distribution is minimized since it is minimized with respect to every value of h.

[2] Here, $\mathbb{E}_{W,H}[(W - \mathbb{E}[W \mid H])^2] = \sum_{w,h} \mathbb{P}[W = w, H = h] \times (w - \mathbb{E}[W \mid H = h])^2$.

4.1 Application and warning

An example of doing this, informally, might be grouping people into buckets according to height, and predicting the weight of a random person in a bucket as the "average" or expectation of the people in the bucket.

The process is a way of approximating the joint distribution of height and weight. Statisticians are careful when doing so to distinguish the sample mean from the real underlying mean (or the conditioned expectation) for example by denoting the sample mean by $\tilde{\mu}$ and the real and unknown mean by μ and being careful about some other terminology. To do this properly is both subtle and interesting but for another time.

In this case, we presume one knows the distribution. This is fine when we are working with a distribution like the geometric or binomial distribution and know the parameters, but with the height and weight example, one may never truly know the joint distribution. On the other hand, this frame is a solid starting point and is mathematically rigorous (given perfect knowledge of the joint distribution) and as the number of samples gets large, the real life situation approaches the full information situation.

5 Linear Regression

5.1 Discussion

We introduced covariance and the correlation coefficient earlier in the course, of which the latter is perhaps the most often reported quantity regarding relationships between two quantities. Even in CS70, we often compute this coefficient for midterm scores and final exam scores to "measure" the consistency of our exams. To be sure, this is the primary measure used in science and the first cut in numerous prediction or estimation problems. In particular, in CS70, "linear regression" on midterm scores typically "explains" around 80% of the variance in on final scores.

This is a mathematical statement about two random variables, midterm score and final score, defined on the sample space of students who take both.

The statement is based on a prediction or estimation method called linear regression, which we will now discuss.

5.2 The mathematics

We consider two random variables, X and Y on some sample space. Recall that for a given value x for X, the best prediction for Y for minimizing mean squared error is simply $\mathbb{E}[Y \mid X = x]$. Note that this prediction could be any function of x as it is defined by the joint distribution.

Say instead, we make predictions using a *linear* function of X. That is, what is the function $\mathscr{L}(X) = mX + b$ that minimizes the mean squared error, defined as $\mathbb{E}\left[(Y - \mathscr{L}(X))^2\right]$? In particular, we wish to derive values for m and b to produce the best linear estimator for Y given a value x for the random variable X.

For simplicity, we will shift the variables to have mean zero. That is, we consider $\bar{X} = X - \mathbb{E}[X]$ and $\bar{Y} = Y - \mathbb{E}[Y]$. Note for a value x for the random variable X, we have $\bar{x} = x - \mathbb{E}[X]$ for the random variable $\bar{X}$.

Furthermore, recall that $\text{Var}(\bar{X}) = \text{Var}(X)$, $\text{Var}(\bar{Y}) = \text{Var}(Y)$, and $\text{cov}(\bar{X}, \bar{Y}) = \text{cov}(X, Y)$.

To estimate $\bar{Y}$ for some value x, we will consider a function $\bar{f}(x) = m\bar{x}$ (that is we assume for now that $b = 0$). We wish to minimize the mean squared error over the entire joint distribution over X and Y, which corresponds to minimizing $\mathbb{E}_{X,Y}[(\bar{Y} - m\bar{X})^2]$.

We can expand this expression as follows:

$$\mathbb{E}\left[(\bar{Y} - m\bar{X})^2\right] = \mathbb{E}\left[\bar{Y}^2\right] - 2\,\mathbb{E}[\bar{X}\bar{Y}] \times m + \mathbb{E}\left[\bar{X}^2\right] \times m^2.$$

We see a quadratic function here in the unknown m of the form $C + Bm + Am^2$ where $A = \mathbb{E}\left[\bar{X}^2\right]$, $B = -2\,\mathbb{E}[\bar{X}\bar{Y}]$, and $C = \mathbb{E}\left[\bar{Y}^2\right]$, which is minimized at $m = -\frac{B}{2A} = \frac{\mathbb{E}[\bar{X}\bar{Y}]}{\mathbb{E}[\bar{X}^2]}$.

We derived this, again, using just our knowledge of a parabola. We assumed that our linear function, $mx + b$, has $b = 0$ or “goes through” the point $(0,0)$ for $\bar{X}$ and $\bar{Y}$. The next subsection justifies this assumption for those interested in the details.

5.3 What about the intercept?

We justify the previous assumption that $b = 0$ for mean zero X and Y by minimizing $\mathbb{E}\left[(\bar{Y} - (m\bar{X} + b))^2\right]$ with respect to the choice of b. Here, we expand to obtain

$$\begin{aligned}
\mathbb{E}\left[(\bar{Y} - (m\bar{X} + b))^2\right] &= \mathbb{E}\left[\bar{Y}^2 - 2\bar{Y}(m\bar{X} + b) + (m\bar{X} + b)^2\right] \\
&= \mathbb{E}\left[\bar{Y}^2\right] - 2m\,\mathbb{E}[\bar{X}\bar{Y}] - 2b\,\mathbb{E}[\bar{Y}] + m^2\,\mathbb{E}\left[\bar{X}^2\right] + 2mb\,\mathbb{E}[\bar{X}] + b^2 \\
&= (\mathbb{E}[\bar{Y}]^2 - 2m\,\mathbb{E}[\bar{X}\bar{Y}] + m^2\,\mathbb{E}\left[\bar{X}^2\right]) + b(-2\,\mathbb{E}[\bar{Y}] + 2m\,\mathbb{E}[\bar{X}]) + b^2
\end{aligned}$$

Taking the derivative with respect to b and setting the result to zero gives the following equation:

$$0 = -2\,\mathbb{E}[\bar{Y}] + 2m\,\mathbb{E}[\bar{X}] + 2b$$

Here, we see that $b = 0$ satisfies the equation, since $\mathbb{E}[\bar{Y}] = \mathbb{E}[\bar{X}] = 0$, which implies that this is where the minimum is.

5.4 Finishing up.

Finally, to get our estimate for Y we substitute back and to estimate $Y - \mathbb{E}[Y]$, use the fact that $\text{cov}(\bar{X},\bar{Y}) = \text{cov}(X,Y)$,$\text{Var}(\bar{X}) = \text{Var}(X)$, $\text{Var}(\bar{Y}) = \text{Var}(Y)$, and solving for Y yields:

$$\mathscr{L}(X) = \frac{\text{cov}(X,Y)}{\text{Var}(X)}(X - \mathbb{E}[X]) + \mathbb{E}[Y].$$

The intuition above is that if X differs from its expectation, then we vary Y from its expectation by an amount proportional to $\text{cov}(X,Y)$. The division by $\text{Var}(X)$ scales the movement in X as well as the dependence of the covariance of X on the typical movement in X.

We can also write the formula as follows to see it as a linear function in the variable X.

$$\mathscr{L}(X) = \frac{\text{cov}(X,Y)}{\text{Var}(X)}X - \frac{\text{cov}(X,Y)}{\text{Var}(X)}\,\mathbb{E}[X] + \mathbb{E}[Y].$$

Exercise. Knowing that $\mathscr{L}(X)$ goes through $(0,0)$ for zero mean random variables X and Y, what point does the $\mathscr{L}(X)$ always go through for general random variables?

We sometimes refer to this estimator as the *linear least squares estimate* of Y given X, which we denote by $\text{LLSE}(Y \mid X)$. Here, we note that this is something that can be calculated from knowing the joint distribution for X and Y; this assumes more than having access to sample points from a distribution, which is often the only access one has to distributions in statistics.

5.5 Linear estimation as explanatory of variance

Armed with our derivation, we make formal the phrase that "X explains some fraction of the variance of Y". In particular, we compare the mean squared error for regression (which uses X) to the variance of Y (which recall is the mean squared error one gets using the estimate $\mathbb{E}[Y]$, which does not use X.) This reduction is the benefit of using a linear function on the value of X to predict Y.

Recall the mean squared error for our problem was this:

$$\mathbb{E}\left[\bar{Y}^2\right] - 2m\mathbb{E}[\bar{X}\bar{Y}] + m^2\mathbb{E}\left[\bar{X}^2\right],$$

and plugging in $m = \frac{\text{cov}(\bar{X},\bar{Y})}{\text{Var}(\bar{X})}$, we obtain the expression:

$$\mathbb{E}\left[\bar{Y}^2\right] - \frac{\text{cov}(\bar{X},\bar{Y})^2}{\text{Var}(\bar{X})}.$$

Dividing by $\text{Var}(\bar{Y}) = \mathbb{E}\left[\bar{Y}^2\right]$, we obtain the expression:

$$1 - (\text{Corr}(\bar{X},\bar{Y}))^2.$$

That is, the correlation coefficient squared is exactly the fraction by which the mean squared error of the linear regression estimator is less than the error in estimating Y using the estimate $\mathbb{E}[Y]$. Thus, the square of the correlation coefficient tells us how much the variance is explained by a linear estimator given X.

6 Statistics

In statistics, one often has some number of samples from a distribution or joint distribution and one wishes to find the line that "best" fits the data where the error is defined as the average squared distance to the line. One basic approach is to assume a uniform distribution over the points themselves and in the limit of large n the above approach works to converge to the LLSE estimator.

To be sure, a Statistician would carefully understand how this approximates the unknown joint distribution. Indeed, the idea of estimating confidence intervals for the coefficients of the regression line as well, though again in the limit the law of large numbers does apply and these values converge.

Statisticians deal with the issue of samples by thinking through the consequences of the sample mean and sample variances and covariances being different for the true ones. Indeed, we have heard exasperation from our statistics colleagues when one confuses the true expectation (typically denoted by μ) with a sample expectation (typically denoted by $\tilde{\mu}$.)

7 Other types of error

The mean square error developed here is quite common in science and applications of statistics. You might even read the phrase X "explains some percentage of the variability" in Y in your local newspaper. This is just the correlation coefficient squared as noted above.

Other notions of error (for example the absolute value of the error) can be more effective for various applications and one might see them in future EECS, Data Science, Statistics or many other departments' curriculums.

Exercise. Prove that the median value of random variable, i.e, x such that $\mathbb{P}[X \geq x] = \mathbb{P}[X \leq x]$, minimizes $\mathbb{E}[|X - x|]$.

8 Prediction from more variables

These days one predicts from many variables; e.g., predicting whether an image is a cat or dog from various features in the image. The above discussion can be generalized to estimating a variable Y from many random variables, $X_1,\dots,X_n$. With a bit of linear algebra, the framework above can be adjusted to this situation. Again, this is an oft-used technique that may be discussed in courses that are in your future.