Math115; Linear algebra and discrete probability review

1. Least squares.

Given n points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ in the (xy)-plane, the least squares line $y = mx + b$ minimizes the total error $E = \sum_{i=1}^{n}(mx_i + b - y_i)^2$. To find this line one should solve with respect to m and b the system of two linear equations

$$\begin{cases} nb + (\sum_{i=0}^{n} x_i)m = \sum_{i=0}^{n} y_i \\ (\sum_{i=0}^{n} x_i)b + (\sum_{i=0}^{n} x_i^2)m = \sum_{i=0}^{n} x_i y_i \end{cases}$$

Most of the coefficients in this system are given by sums like $\sum_{i=0}^{n} x_i, \sum_{i=0}^{n} x_i y_i$, etc. It is easier to compute them separately (since x_i's and y_i's are given numbers) before writing down the linear system.

2. Matrix addition and multiplication.

- Size constraints: $(m\times n)+(m\times n) \mapsto m\times n$, $(m\times p)\cdot(p\times n) \mapsto m\times n$.
- Distributivity and associativity: $(A+B)C = AC+BC$, $C(A+B) = CA+CB$, $A(BC) = (AB)C$ when defined.
- It is possible that $AB \neq BA$ or $AB = 0$ for non-zero 2×2 matrices. For example, $AB = 0 \neq BA$ where
$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$$
- Any linear system may be written as $A\vec{x} = \vec{b}$; the system is homogeneous when $\vec{b} = 0$.

3. Gauss-Jordan elimination.

Elimination preserves the solution set of a system of linear equations. These are the main cases when one should apply elimination:

- To solve a linear system $A\vec{x} = \vec{b}$ apply Gauss-Jordan to $(A|b)$.
- To find A^{-1} apply Gauss-Jordan to $(A|I)$.
- More generally, to solve a matrix equation $AX = B$ apply Gauss-Jordan to $(A|B)$.

Example. If A is an unknown invertible matrix with known A^2 and A^3, then A solves the matrix equation $A^2X = A^3$. Thus, X may be found by applying Gauss elimination to $(A^2|A^3)$ (i.e. $(A^2|A^3)$ is taken to $(I|A)$ by Gauss elimination).

A homogeneous $n\times n$ system $A\vec{x} = 0$ has a non-zero solution if and only if applying Gauss elimination to A we obtain a matrix with a zero row (note also that the latter happens if and only if A has no inverse).

4. Inverse matrix.

- If $AB = I$ for $n\times n$ matrices, then $BA = I$ and $B = A^{-1}, A = B^{-1}$.
- Properties: $(A^{-1})^{-1} = A$,$(AB)^{-1} = B^{-1}A^{-1}$.
- The solution of a linear system $A\vec{x} = \vec{b}$ with invertible A is $\vec{x} = A^{-1}\vec{b}$.

5. Input-output models.

- The demand matrix D is a column whose entries are demands for various commodities (in dollars).
- The production matrix X is a column whose entries are production quantities (in dollars). Usually it is the unknown.
- The input-output matrix A is a matrix whose i-th column entries are the quantities of various commodities (in dollars) used to produce the i-th commodity.
- The main equation is $D = X - AX = (I - A)X$. It may be solved with respect to X by Gauss-Jordan elimination.
- If the model is *open*, i.e. the demand is not zero, then usually $I - A$ is invertible and then $X = (I - A)^{-1}D$.
- If the model is *closed*, i.e. the demand is zero, then A is a *right stochastic* matrix, i.e. the sum of entries in each its column is 1.
- It may be deduced that $I - A$ is not invertible. In particular, the homogeneous system of linear equations $(I - A)X = D = 0$ has at least one-parametric family of solutions (which can be found by elimination).

6. Probability: basic notions.

If A, B are subsets of a set S, then AB denotes their intersection and $A \cup B$ denotes their union. If AB is empty, then we say that A and B are *disjoint*, call their union the *disjoint union* and denote it $A \sqcup B$.

Definition. A *sample space* S is a space provided with a function pr attaching to its subsets A non-negative numbers $\mathrm{pr}(A)$ and such that $\mathrm{pr}(S) = 1$ and $\mathrm{pr}(A \sqcup B) = \mathrm{pr}(A) + \mathrm{pr}(B)$ for each pair of disjoint subsets $A, B \subset S$.

- A sample space S is usually interpreted as the set of all possible *outcomes* of an experiment.
- Subsets of a sample set are called *events*. Thus, we can talk about probability of events.

Examples:

(i) If one tosses a coin n-times then the sample space consists of 2^n possible outcomes (sequences of heads and tails of length n). Examples of events in such space are: the first three times one gets heads; there are more tails than heads; the first and the last toss give the same result, etc. All outcomes have probability $\frac{1}{2^n}$.

(ii) If one tosses dices n times then the sample space consists of 6^n possible outcomes. Examples of events are as follows: all results are even; the sum of n results is $2n + 3$, etc. All outcomes have probability $\frac{1}{6^n}$.

Definition. A probability space is called *simple* if all outcomes occur with the same probability.

Fact. A simple space is necessarily finite, say of size m. Then for any event A of size k one has $\mathrm{pr}(A) = \frac{k}{m}$. In particular, the probability of each outcome is $\frac{1}{m}$.

7. Simple spaces and counting methods.

- To find the probability of an event A in a simple sample space S one just has to find the sizes of A and S. In many cases it may be done by counting or combinatorial methods.
- If an experiment consists of m stages and we can get n_1 outcomes on the first stage, n_2 outcomes on the second one, etc. Then the total number of outcomes is the product $n_1 \cdot n_2 \cdot n_3 \cdot \dots \cdot n_m$.
- In particular, there are $n(n-1)\dots(n-k+1)$ ways to choose an ordered k-tuple of objects from a total set of n objects. The number $P_{n,k} = n(n-1)\dots(n-k+1) = \frac{n!}{(n-k)!}$ is called the number of permutations of n elements taken k at a time.
- Taking $k = n$ we see that there are exactly $n!$ ways to order a set of n elements (it is the number of permutations of a set of n elements).
- The number $C_{n,k} = \binom{n}{k}$ of k-element subsets of a set of n elements is called the number of combinations of n elements taken k at a time, or the binomial coefficient. It equals to $P_{n,k}$ divided by the number of possible orders on a k-element set. Thus, $\binom{n}{k} = \frac{P_{n,k}}{k!} = \frac{n!}{k!(n-k)!}$ (where we agree that $0! = 1$).

Here are some properties of binomial coefficients.

- $\binom{n}{0} = \binom{n}{n} = 1$
- $\binom{n}{1} = \binom{n}{n-1} = n$
- $\binom{n}{2} = \binom{n}{n-2} = \frac{n(n-1)}{2}$
- $\binom{n}{k} = \binom{n}{n-k}$
- $\binom{n+1}{k+1} = \binom{n}{k+1} + \binom{n}{k}$
- $\binom{n}{0} + \binom{n}{1} + \dots + \binom{n}{n} = 2^n$

Try to explain the last three properties combinatorially! The following theorem is the famous binomial theorem of Newton.

Theorem. An n-th power of a sum may be expanded as

$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i} =$$

$$x^n + nx^{n-1}y + \frac{n(n-1)}{2}x^{n-2}y^2 + \binom{n}{3}x^{n-3}y^3 + \dots + \frac{n(n-1)}{2}x^2y^{n-2} + nxy^{n-1} + y^n$$

Substitute $x = y = 1$ in this identity. Which identity involving binomial numbers do you get?

8. Multinomial numbers.

Given a set S of size n, a binomial number $\binom{n}{k}$ may be interpreted as the number of divisions of S to two subsets S_1 and S_2 of sizes k and $n-k$. More generally, for any sequence of positive natural numbers $n_1, n_2, \dots, n_m$ with the total sum equal to n one defines a multinomial number $\binom{n}{n_1, n_2 \dots n_m}$ as the number of ways to divide S to the sets $S_1, S_2 \dots S_m$ of sizes $n_1, n_2, \dots, n_m$. In particular, $\binom{n}{k, n-k} = \binom{n}{k}$.

Fact. $\binom{n}{n_1, n_2, \ldots, n_m} = \binom{n}{n_1} \cdot \binom{n-n_1}{n_2} \cdot \binom{n-n_1-n_2}{n_3} \cdots = \frac{n!}{n_1! \cdot n_2! \ldots n_m!}$

There is a multinomial formula for $(x_1 + x_2 + \ldots x_m)^n$ which involves multinomial coefficients, but usually it is very messy and rather useless.

9. Probability of a union.

If A is a disjoint union of $A_1, A_2, \ldots, A_m$, then $\mathrm{pr}(A)$ equals to the sum of all $\mathrm{pr}(A_i)$ But if the sets are not disjoint, then the formula for $\mathrm{pr}(A)$ in terms of A_i's becomes more involved and requires to know the probabilities of the intersections. We give here only the formulas for $m \leq 3$.

If $A = A_1 \cup A_2$, then

$$\mathrm{pr}(A) = \mathrm{pr}(A_1) + \mathrm{pr}(A_2) - \mathrm{pr}(A_1 A_2)$$

If $A = A_1 \cup A_2 \cup A_3$, then

$$\mathrm{pr}(A) = \mathrm{pr}(A_1) + \mathrm{pr}(A_2) + \mathrm{pr}(A_3) - \mathrm{pr}(A_1 A_2) - \mathrm{pr}(A_2 A_3) - \mathrm{pr}(A_1 A_3) + \mathrm{pr}(A_1 A_2 A_3)$$

10. Conditional probability.

- If A and B are two events in a sample space S and $\mathrm{pr}(B) \neq 0$, then the conditional probability $\mathrm{pr}(A|B)$ of A given B is defined as

$$\mathrm{pr}(A|B) = \frac{\mathrm{pr}(AB)}{\mathrm{pr}(B)}$$

- Intuitively it means that we cut everything by intersecting with B which replaces the old sample space S. We have to divide the probabilities by $\mathrm{pr}(B)$ to achieve that $\mathrm{pr}(B|B) = 1$.
- We say that A and B are *independent* if $\mathrm{pr}(A|B) = \mathrm{pr}(A)$ (i.e. the probability of A does not change when we know that B happened). Equivalently, A and B are independent if and only if $\mathrm{pr}(AB) = \mathrm{pr}(A)\mathrm{pr}(B)$.

11. Bayes' theorem.

If a sample set S is a disjoint union of subsets, say $S = S_1 \sqcup S_2 \sqcup \cdots \sqcup S_m$ and A is an event, then the *total probability law* states that

$$\mathrm{pr}(A) = \mathrm{pr}(AS_1) + \mathrm{pr}(AS_2) + \cdots + \mathrm{pr}(AS_m) =$$

$$\mathrm{pr}(A|S_1)\mathrm{pr}(S_1) + \mathrm{pr}(A|S_2)\mathrm{pr}(S_2) + \cdots + \mathrm{pr}(A|S_m)\mathrm{pr}(S_m)$$

Since $\mathrm{pr}(S_i|A) = \mathrm{pr}(AS_i)/\mathrm{pr}(A)$ for any $1 \leq i \leq m$, one can deduce from the total probability law the following theorem discovered by Bayes

$$\mathrm{pr}(S_i|A) = \frac{\mathrm{pr}(A|S_i)\mathrm{pr}(S_i)}{\mathrm{pr}(A|S_1)\mathrm{pr}(S_1) + \mathrm{pr}(A|S_2)\mathrm{pr}(S_2) + \cdots + \mathrm{pr}(A|S_m)\mathrm{pr}(S_m)}$$

Notice that Bayes theorem allows to compute the conditional probability $\mathrm{pr}(S_i|A)$ in terms of the conditional probabilities $\mathrm{pr}(A|S_i)$. Often $\mathrm{pr}(S_i|A)$ can be interpreted as the posterior probability of S_i given that A happened. So, Bayes' theorem is very useful in computing posterior probabilities.

12. Markov's chains.

Instead of discussing the general theory we consider an example. Assume that A, B, C play catch, A throws to B and C with probability $\frac{1}{2}$, B throws to A with probability $\frac{1}{3}$ and throws to C with probability $\frac{2}{3}$ and C throws to A and B with probability $\frac{1}{3}$ and forgets to throw a ball in one third of all cases. What is the probability that C will have the ball in a long term run?

We form a transition matrix

$$P = \begin{pmatrix} \mathrm{pr}(A \to A) & \mathrm{pr}(B \to A) & \mathrm{pr}(C \to A) \\ \mathrm{pr}(A \to B) & \mathrm{pr}(B \to B) & \mathrm{pr}(C \to B) \\ \mathrm{pr}(A \to C) & \mathrm{pr}(B \to C) & \mathrm{pr}(C \to C) \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{3} \\ \frac{1}{2} & \frac{2}{3} & \frac{1}{3} \end{pmatrix}$$

Notice that P is a right stochastic matrix similarly to the case of closed input-output models, i.e. the sum of the entries in each column is equal to 1. **Our choice is opposite to the choice of the text**, where one produces a left stochastic matrix (with sums in rows equal to 1).

If a vector

$$\vec{p}_n = \begin{pmatrix} \mathrm{pr}(A, n) \\ \mathrm{pr}(B, n) \\ \mathrm{pr}(C, n) \end{pmatrix}$$

consists of the probabilities that A, B or C possess the ball after n stages, then we can compute similar probabilities in the future by the transition formula $\vec{p}_{n+1} = P\vec{p}_n$, or more generally

$$\vec{p}_{n+k} = P^k \vec{p}_n$$

Finally, if a vector

$$\vec{p} = \begin{pmatrix} \mathrm{pr}(A) \\ \mathrm{pr}(B) \\ \mathrm{pr}(C) \end{pmatrix}$$

consists of the probabilities that A, B or C possess the ball in a long term run, then it satisfies the matrix equation $\vec{p} = P\vec{p}$, or equivalently

$$(I - P)\vec{p} = 0$$

Notice that this equation is absolutely similar to the main equation on closed input-output models!

Thus, in our case we have to solve the matrix equation

$$\begin{pmatrix} 1 & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{2} & 1 & -\frac{1}{3} \\ -\frac{1}{2} & -\frac{2}{3} & \frac{2}{3} \end{pmatrix} \vec{p} = 0$$

Applying Gauss elimination to the augmented matrix

$$\left(\begin{array}{ccc|c} 1 & -\frac{1}{3} & -\frac{1}{3} & 0 \\ -\frac{1}{2} & 1 & -\frac{1}{3} & 0 \\ -\frac{1}{2} & -\frac{2}{3} & \frac{2}{3} & 0 \end{array}\right)$$

(see below) one finds a one-parametric solution $x_3 = t, x_2 = \frac{3}{5}t, x_1 = \frac{8}{15}t$. Since x_1, x_2 and x_3 are the probabilities that A, B and C, respectively, has the ball in a long term run, we have also $x_1 + x_2 + x_3 = 1$. Hence $\frac{32}{15}t = 1$, $x_3 = t = \frac{15}{32}$, and we see that C will have the ball slightly less than one half of the whole time.

13. An example of applying Gauss elimination.

We will solve the matrix equation from the previous section (actually, it is a homogeneous system of linear equations).

$$\left(\begin{array}{rrr|r} 1 & -\frac{1}{3} & -\frac{1}{3} & 0 \\ -\frac{1}{2} & 1 & -\frac{1}{3} & 0 \\ -\frac{1}{2} & -\frac{2}{3} & \frac{2}{3} & 0 \end{array}\right) \xrightarrow{6R_1;\ 6R_2;\ 6R_3} \left(\begin{array}{rrr|r} 3 & -1 & 1 & 0 \\ -3 & 6 & -2 & 0 \\ -3 & -4 & 4 & 0 \end{array}\right) \xrightarrow{R_2+R_1;\ R_3+R_1}$$

$$\left(\begin{array}{rrr|r} 3 & -1 & 1 & 0 \\ 0 & 5 & -3 & 0 \\ 0 & -5 & 3 & 0 \end{array}\right) \xrightarrow{R_3+R_2} \left(\begin{array}{rrr|r} 3 & -1 & 1 & 0 \\ 0 & 5 & -3 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

The last system implies that x_3 is a parameter, say $x_3 = t$. From the second row we get $5x_2 = 3x_3$, hence $x_2 = \frac{3}{5}t$. Finally, the first row implies that $3x_1 = x_2 + x_2 = \frac{3}{5}t + t = \frac{8}{5}t$, hence $x_1 = \frac{8}{15}t$.