COS 424 Lecture Notes

Lecturer: L. Bottou
Scribes: J. Valentino & R. Misener

February 18, 2010

1 Administrivia

- Office hours are on an appointment basis. Additionally, L. Bottou is available immediately after class to discuss any questions.
- The goal of this and the next lecture (Thursday, February 18) is to give an introduction to probability and identify the difficult parts. Probability is more difficult than it looks, so L. Bottou wants us to have a solid foundation and a clear understanding of where the difficulties are.
- This lecture also contains a brief introduction to linear algebra because students asked about solving linear systems after a previous lecture.

2 Linear Systems of Equations

Suppose we have a vector of unknowns $\mathbf{x}$, parameter matrix $\mathbf{A}$, and parameter vector $\mathbf{b}$ with $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$. In practice (and throughout this course), we will use existing software (*i.e.*, BLAS and LAPACK) for solving systems of linear equations. However, L. Bottou wants to show how the algorithms work (see the Numerically Solving $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$ section in the *Linear Algebra Review for COS 424* handout).

The two commonly-used linear algebra packages, BLAS and LAPACK (which uses BLAS to solve equations) are old but of very high quality. They have been worked out to the level of very minute details. Matlab and R both internally use BLAS and LAPACK. Intel has a version of BLAS that is optimized for individual processors.

2.1 What NOT to do and why

- **Invert $\mathbf{A}$** (*i.e.*, directly compute $\mathbf{A}^{-1}$). Inverting $\mathbf{A}$ effectively means solving n equations equivalent to $\mathbf{A} \cdot \mathbf{u} = \mathrm{e}_i$ such that $i \in n$ where n is the number of columns in $\mathbf{A}$ and e_i is a column vector of all zeros except for a 1 in position i. Solving these n equations implies that inverting $\mathbf{A}$ takes n times the necessary work.
- **Use Cramer's Rule.** Cramer's Rule calculates each $\mathbf{x}_i$ using a ratio of determinants:

$$\mathbf{x}_i = \frac{\det(\mathbf{A}_i)}{\det(\mathbf{A})} \quad \forall\, i \in n$$

  where $\mathbf{A}_i$ is the matrix formed by replacing column i of $\mathbf{A}$ by column vector $\mathbf{b}$. Because calculating a determinant is almost as costly as calculating an inverse, using Cramer's Rule requires approximately

$n+1$ times the work as inverting a matrix or O(n^2) the necessary work. Cramer's Rule is taught in grade school because it's easy to understand, but it's a computational catastrophe.

2.2 Interesting Matrices

- **Triangular Matrix** Solving $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$ for a triangular matrix is easy. For $n = 3$:

$$\begin{aligned} a_{11} \cdot x_1 + \quad a_{12} \cdot x_2 + \quad a_{13} \cdot x_3 &= b_1 \\ a_{22} \cdot x_2 + \quad a_{23} \cdot x_3 &= b_2 \\ a_{33} \cdot x_3 &= b_3 \end{aligned}$$

  Take the last equation, find x_3 right away, back-substitute it into the middle equation, etc. Each step has a single unknown, so it's easy to solve. The entire process is computationally equivalent to multiplying a matrix and a vector.

- **Orthogonal Matrix** An orthogonal matrix is composed of orthogonal columns with the unit norm. Orthogonal matrices have the property that $\mathbf{A}^T = \mathbf{A}^{-1}$ because each term of $\mathbf{A}^T \cdot \mathbf{A}$ is the dot product of one column and another. If the columns are different than one another, their dot product will be zero. If they are the same, their dot product will be one.

2.3 Decomposition Approaches

BLAS and LAPACK both use decomposition approaches to solve linear systems of equations. Decomposition approaches work by *decomposing* $\mathbf{A}$ into triangular and orthogonal matrices. We should not be programming these things ourselves – this introduction is just to show what is going on *under the hood*.

- **QR** This approach re-writes square invertible matrix $\mathbf{A}$ as $\mathbf{A} = \mathbf{Q} \cdot \mathbf{R}$ where $\mathbf{Q}$ is orthogonal and $\mathbf{R}$ is triangular. After decomposition, the matrix is simple to solve because:

$$\begin{aligned} \mathbf{A} \cdot \mathbf{x} &= \mathbf{b} \\ \mathbf{Q} \cdot \mathbf{R} \cdot \mathbf{x} &= \mathbf{b} \\ \mathbf{R} \cdot \mathbf{x} &= \mathbf{Q}^T \cdot \mathbf{b} \end{aligned}$$

  As described in the previous section, the final line is easy to solve.

- **LU** This approach re-writes square invertible matrix $\mathbf{A}$ as $\mathbf{A} = \mathbf{L} \cdot \mathbf{U}$ where $\mathbf{L}$ is a a lower triangular matrix $\mathbf{U}$ is upper triangular:

$$\begin{aligned} \mathbf{A} \cdot \mathbf{x} &= \mathbf{b} \\ \mathbf{L} \cdot \mathbf{U} \cdot \mathbf{x} &= \mathbf{b} \\ \mathbf{U} \cdot \mathbf{x} &= \text{something} \end{aligned}$$

  This method solves for $\mathbf{U} \cdot \mathbf{x}$ and does another back substitution to find $\mathbf{x}$

- **SVD** To be discussed later when we have more values to discuss.

There are many algorithms to perform decomposition, but here's an example of **QR** decomposition using the Gram-Schmidt process. Gram-Schmidt is *not* the computationally best algorithm, but it's an elegant one.

The goal is to get pairwise orthogonal $\mathbf{Q}$ with unit norm columns. Consider matrix $\mathbf{A} = [\mathbf{u}_1,\ \mathbf{u}_2,\ \mathbf{u}_3]$ with $n = 3$. We want to build an orthonormal basis $\mathbf{Q} = [\mathbf{q}_1,\ \mathbf{q}_2,\ \mathbf{q}_3]$ where the $\mathbf{q}_i$ are column vectors that spans the same space.

First we take $\mathbf{u}_1$ and normalize it:

$$\mathbf{q}_1 = \frac{\mathbf{u}_1}{\|\mathbf{u}_1\|} \Longrightarrow \mathbf{u}_1 = r_{11} \cdot \mathbf{q}_1$$

For the second column, subtract the residual (to make the two columns orthogonal) and normalize the result:

$$\mathbf{x} = \mathbf{u}_2 - (\mathbf{x} \cdot \mathbf{u}_1) \cdot \mathbf{u}_1 \Longrightarrow \mathbf{q}_2 = \frac{\mathbf{x}}{\|\mathbf{x}\|} \Longrightarrow \mathbf{u}_1 = r_{21} \cdot \mathbf{q}_1 + r_{22} \cdot \mathbf{q}_2$$

Keep repeating this process of subracting off the residuals from the previous columns and normalizing the result. This generates the appropriate $\mathbf{Q}$ and $\mathbf{R}$ for **QR** decomposition.

One possible numerical difficulty is that one of the $\mathbf{u}_i$ is almost in the subspace spanned by the previously-generated columns. This will make $\mathbf{x}$ very small and introduce numerical error. There are ways around this. For example, you could pre-select the order of the columns such that the next column selected leads to a big difference. Fancy algorithms like this are integrated into BLAS and LAPACK, so don't sweat the details.

3 Probability

We all have been exposed to informal probabilities, but probability is fairly subtle. Probability is part of the common language which sometimes mistakenly leads us to believe that we know what is going on, but we may not. The idea of probability is not necessarily easy or well understood. Pascal made breakthroughs in the 17^{th} century, but complete and clear axioms of probability were only developed in the 20^{th} century.

We will review probability so that L. Bottou can give some perspective on what the difficult problems are, but he does not expect us to become deep experts in Borel algebra, measure theory, etc. (just know it exists).

3.1 Discrete Probabilities

We consider discrete probabilities with finite sets and probabilities in finite spaces beacuse they're relatively reasonable to deal with and resemble real life situations. Difficulties typically come from discussing continuous probabilities, which are the limit of what we can observe and often applied for mathematical convenience.

As an example, We can assume we are dealing with a random process that depends on k random coin tosses or randomly picking an atom from a space. We can describe an event as a particular sequence of coin tosses (*e.g.*, HTHH).

Set Ω is the space of all possibilities and each element within the set is a sequence of coin tosses, dice rolls, etc. Each possible event $\omega \in \Omega$ is inside the space.

To switch into the probability space, convert each atom or dice roll into a measure (*e.g.*, count of occurances or measurement of mass). The measure $m(\omega)$ is a real number. The probability of an event is its measure divided by the sum over Ω of the measure of everything else:

$$p(\omega) = \frac{m(\omega)}{\sum_{\omega \in \Omega} m(\omega)}$$

Suppose we are dealing with atoms and we want to address more complicated problems. We want *events* $A \subset \Omega$ so that we can consider lots of atoms. The probabilty of an event A occuring is the sum of all of the atoms that belong in the event times the probability of the atom:

$$P(A) = \sum_{\omega \in A} p(w)$$

In the language of probabilities, a random process picks an atom and we test to determine if it is in an event. The language of set theory can help make things more simple. Some random process will pick an atom. Some random event occurs if the atom that is picked belongs to that event. We say that:

- Events A AND B occur simultaneously if the atom belongs to the intersection of these events.
- Event A OR B occurs if the atom belongs to the union of these events (inclusive or).
- Either A OR B occurs if the atom belongs to the union and but not the intersection of the events (exclusive or).

This a similarity between the event language and the set theory language allows us to use set theory to ground the probabilities. There are two *essential properties*:

$$\begin{aligned} P(\Omega) &= 1 \\ A \cap B &= \emptyset \Longrightarrow P(A \cup B) = P(A) + P(B) \end{aligned}$$

From these two *essential properties* we can find the three following *derived properties*. First:

$$P(A^C) = 1 - P(A)$$

To see this, imagine A and B as a venn diagram in Ω. Second:

$$P(\emptyset) = 0$$

by combining $P(\Omega) = 1$ and the first derived property. Finally:

$$P(A \cup B) = P(A) + P(B)P(A \cap B)$$

3.2 Random Variables

A *random variable* X is a function that maps Ω to X and ω to X(ω). For example, if you roll dice, X could be the sum of the rolls of the dice. In this case, the output space is also a probability space. For example, if you have B $\subset$ X, we can define the probability:

$$P_X(B) = P(X \in B) = P\left\{\omega : x(\omega) \in B\right\}$$

This equation allows us to transport a predicate involving random variables to a subset on the space of atoms:

$$P(X < x) = P\{\omega : x(\omega) < x\}$$

When we defined a random variable we transported the probabilities into another set, which is output of the random variable itself. Another method of transport is to use *conditional probabilities*.

3.3 Conditional Probabilities

Suppose we know that event A occurs, so we are now no longer interested in all of Ω, only in the atoms that are within event A. We now want probabilities that are defined on A only. Recalling that:

$$P(\omega) = \frac{m(\omega)}{\sum\limits_{\omega \in \Omega} m(\omega)}$$

We now define:

$$P_A(\omega) = \frac{m(\omega)}{\sum\limits_{\omega \in A} m(\omega)}$$

Now if we have another subset B, the probability that B occurs given A is:

$$P(B|A) = \sum_{\omega \in A \cap B} P_A(\omega) = \frac{P(A \cap B)}{P(A)} = \frac{P(A,\, B)}{P(A)}$$

Alternatively, $P(A,\, B) = P(A) \cdot P(B|A)$ and $P(A,\, B) = P(B) \cdot P(A|B)$. Putting this together we get Bayes' theorem:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Bayes' theorem is a powerful tool because it is inversion of evidence. Suppose we roll a die four times $(x_1,\, x_2,\, x_3,\, x_4)$ and we know that the sum of the counts is 6. We want to find the probability of having $x_2 = 3$. This, $P(x_2 = 3|s = 6)$, is hard to calculate directly, but we can transform this into a simple combinatorial problem with Bayes' theorem because $P(s = 6|x_2 = 3)$ is simple to calculate.

3.4 Chain Theorem

Suppose we have events $A_1,\, A_2,\, \ldots,\, A_n$. The problem of all of them occuring is:

$$P(A_1,\, A_2,\, \ldots,\, A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1,\, A_2) \cdot \ldots \cdot P(A_n|A_1,\, A2,\, \ldots,\, A_{n-1})$$

This is a useful result because it allows a complicated distribution to be split into manageable little pieces.

3.5 Marginalization

Suppose we know that $P(A|X = x) \quad \forall\ x$. Using the essential probability rules and the sum of the joint events, we can say:

$$P(A) = \sum_{x \in X} P(A, X = x) = \sum_{x \in X} P(A|X = x) * P(X = x)$$

Therefore, if we know the conditional probabilities, we can reconstruct the probability of the base event.

3.6 Monty Hall Problem

We will take the standard Monty Hall Problem and study it in deeper detail than we usually do. The goal here is to show that probabilities are more than just frequencies. See the review on probabilities for the math (the Monty Hall Problem section). Probabilities are a great tool because they allow both theoretical modeling and connecting to the real data.

(See lecture notes here.)

If you look at statisticians, they are split into two camps those who care about dealing with frequency and those who look at probabilities as a tool for reasoning and therefore pay much less attention to repetitions for events (Bayesians). Both have some truth – we should use both where appropriate.

3.7 Expectation

Suppose that we have a random variable X that takes value in a space where we can perform operatins such as multiplication by a scalar or addition. The mathematical expectation of X is:

$$E[x] = \sum_{x \in X} x \cdot P(X = x)$$

Recall that we are still in a discrete space, so x will be finite and $P(X = x)$ will be nonzero. If we were dealing with the continuous space, then we would have to discuss if $P(X = x)$ in the limit.

Expectations have interesting properties. The indicator function is:

$$\mathbb{1}_{[A]} = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}$$

and it has the property:

$$E(\mathbb{1}_{[A]}) = P(A)$$

Another thing that derives from the definition is that the expectation is linear in X:

$$E(X + Y) = E(X) + E(Y)$$

Additionally:

$$E(\alpha \cdot X) = \alpha \cdot E(X)$$

and:

$$E(X \cdot Y) = E(X) \cdot E(Y)$$

if X and Y are independent. Intuitively, the expectation is the average value of X. Assume you toss coins and if you get heads you get \$1 and if you get tails you get \$0. The expectation of earnings will be 50 cents per toss.

3.8 Variance and Co-Variance

Use the variance to find how close X is to the expectation:

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

We can perform these operations because X is a random variable and E(X) is a scalar. The scalar can be moved ouside of the expectation because of property the third property of expectations. We can generalize variance to multiple variables and find the covariance:

$$Cov(X, Y) = E\left[(X - E(X)) \cdot (Y - E(Y))\right] = E(XY) - E(X)E(Y)$$

If X and Y are independent then covariance will be zero and they are de-correlated.

3.9 Independence and Dependence

We would like to capture the fact that events can occur independently from one another. If $P(A|B) = P(A)$ then A and B are pairwise independent and knowing B tells us nothing about A. This is equivalent to saying $P(A,\ B) = P(A) \cdot P(B)$. Two events are independent if the probability of the joint set is the same as the product of the independent probabilities themselves.

Discussing the case where three events are independent is not the same as saying that they are pairwise independent. Suppose we toss two coins and consider three events: $C_1 = H$, $C_2 = H$, $C_1 = C_2$. The three events are pairwise independent, but are not independent, so the definition in this case is more complicated.