AIMS | African Institute for Mathematical Sciences
RWANDA

AMMI Review sessions

Deep Learning (1)

Applied Math

10 Feb 2020

Linear Algebra

- Linear combination:
- Multiplying each vector v$^{(i)}$ by a corresponding scalar coefficient and adding the results

$$\sum_i c_i \boldsymbol{v}^{(i)}.$$

Linear Algebra

- **Linear combination:**
- Multiplying each vector $v^{(i)}$ by a corresponding scalar coefficient and adding the results

$$\sum_i c_i \boldsymbol{v}^{(i)}.$$

- The span of a set of vectors is the set of all points obtainable by linear combination of the original vectors.

Linear Algebra

- **Linear combination:**
- Multiplying each vector $v^{(i)}$ by a corresponding scalar coefficient and adding the results

$$\sum_i c_i \boldsymbol{v}^{(i)}.$$

- The span of a set of vectors is the set of all points obtainable by linear combination of the original vectors.

- Determining whether Ax = b has a solution thus amounts to testing whether b is in the **span** of the columns of A. This particular span is known as the column space or the **range** of A.

Linear Algebra

- Linear combination:

- In order for the system Ax = b to have a solution for all values of b ∈ R^m, we therefore require that the column space of A be all of R^m, this implies A must have at least m columns i.e., n ≥ m.

- Having n ≥ m is only a necessary condition for every point to have a solution. It is not a sufficient condition, because it is possible for some of the columns to be redundant.

- Formally, this kind of redundancy is known **as linear dependence**. A set of vectors is **linearly independent** if no vector in the set is a linear combination of the other vectors.

Linear Algebra

- Linear combination:
- Having n ≥ m is only a necessary condition for every point to have a solution. It is not a sufficient condition, because it is possible for some of the columns to be redundant.

- Formally, this kind of redundancy is known as linear dependence. A set of vectors is linearly independent if no vector in the set is a linear combination of the other vectors.

Linear Algebra

- **Linear combination:**
- The matrix must contain at least one set of m linearly independent columns. This condition is both necessary and sufficient for equation Ax = b
- In order for the matrix to have an inverse, we additionally need to ensure that equation Ax=b has at most one solution for each value of b

Linear Algebra

- **Linear combination:**
- The matrix must contain at least one set of m linearly independent columns. This condition is both necessary and sufficient for equation Ax = b

- In order for the matrix to have an **inverse**, we additionally need to ensure that equation Ax=b has at most one solution for each value of b, **this means that the matrix must be square**

- this means that the matrix must be square, that is, we require that m = n and that all of the columns must be linearly independent. A square matrix with linearly dependent columns is known as singular.

Linear Algebra

- **Linear combination:**

- The matrix must contain at least one set of m linearly independent columns. This condition is both necessary and sufficient for equation Ax = b

- In order for the matrix to have an **inverse**, we additionally need to ensure that equation Ax=b has at most one solution for each value of b, **this means that the matrix must be square**

- this means that the matrix must be square, that is, we require that m = n and that all of the columns must be linearly independent. A square matrix with linearly dependent columns is known as singular.

- If A is not square or is square but singular, it can still be possible to solve the equation. However, we can not use the method of matrix inversion to find the solution

Linear Algebra

Norms:

- we usually measure the size of vectors using a function called a norm (mapping vectors to non-negative values) Formally, the L norm is given by:

$$||\boldsymbol{x}||_p = \left(\sum_{i} |x_i|^p \right)^{\frac{1}{p}}$$

Linear Algebra

Norms:

- we usually measure the size of vectors using a function called a norm (mapping vectors to non-negative values) Formally, the L norm is given by:

$$||\boldsymbol{x}||_p = \left(\sum_i |x_i|^p \right)^{\frac{1}{p}}$$

- **The L2 norm**, with p = 2, is known as the Euclidean norm. It is simply the Euclidean distance from the origin to the point identified by x

- The L2 norm is used so frequently in machine learning that it is often denoted simply as ||x||,

The squared L2 norm is more convenient to work with mathematically and computationally than the L2 norm itself.

Linear Algebra

Norms:

- In many contexts, the squared L2 norm may be undesirable because it increases very slowly near the origin.

- **L1** is preferred when it is important to discriminate between elements that are exactly zero and elements that are small but nonzero.

$$||\boldsymbol{x}||_1 = \sum_i |x_i|.$$

- Sometimes we may also wish to measure **the size of a matrix**. In the context of deep learning, the most common way use Frobenius norm:

$$||A||_F = \sqrt{\sum_{i,j} A_{i,j}^2},$$

which is analogous to the L2 norm of a vector.

Linear Algebra

Eigendecomposition:

- Many mathematical objects can be understood better by breaking them into constituent parts

- In eigen-decomposition, in which we decompose a matrix into a set of eigenvectors and eigenvalues.

$$Av = \lambda v.$$

- If v is an eigenvector of A, then so is any rescaled vector sv for $s \in R$, s != 0. Moreover, sv still has the same eigenvalue. For this reason, we usually only look for unit eigenvectors.

- The eigendecomposition of A is then given by

$$A = V\text{diag}(\lambda)V^{-1}.$$

Linear Algebra

Eigendecomposition:

- The matrix is singular if and only if any of the eigenvalues are zero.

- The eigendecomposition of a **real symmetric matrix** can also be used to optimize quadratic expressions of the form f(x) = x Ax subject to ||x||2 = 1. Whenever x is equal to an eigenvector of A, f takes on the value of the corresponding eigenvalue.

Linear Algebra

Eigendecomposition:

- The matrix is singular if and only if any of the eigenvalues are zero.

- The eigen-decomposition of a **real symmetric matrix** can also be used to optimize quadratic expressions of the form f(x) = x Ax subject to ||x||2 = 1. Whenever x is equal to an eigenvector of A, f takes on the value of the corresponding eigenvalue.

- A matrix whose eigenvalues are all positive is called **positive definite**. A matrix whose eigenvalues are all positive or zero-valued is called **positive semi-definite.**

- Positive semidefinite matrices are interesting because they guarantee that ∀x,

$$x^T A x \geq 0.$$

- Positive definite matrices additionally guarantee that

$$x^T A x = 0 \Rightarrow x = 0.$$

Linear Algebra

Eigendecomposition:

Singular Value decomposition:

- Every real matrix has a singular value decomposition, but the same is not true of the eigenvalue decomposition. For example, if a matrix is not square, the eigendecomposition is not defined

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$$

- U and V are both defined to be orthogonal matrices, while D is diagonal

Linear Algebra

Eigendecomposition:

Singular Value decomposition:

- Every real matrix has a singular value decomposition, but the same is not true of the eigenvalue decomposition. For example, if a matrix is not square, the eigendecomposition is not defined

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$$

- U and V are both defined to be orthogonal matrices, while D is diagonal

- The elements along the diagonal of D are known as **the singular values of the matrix A**. The columns of U are known as the **left-singular** vectors. The columns of V are known as as the **right-singular** vectors.

- Perhaps the most useful feature of the SVD is that we can use it to partially **generalize matrix inversion** to **non-square** matrices

Linear Algebra

Singular Value decomposition:

- Every real matrix has a singular value decomposition, but the same is not true of the eigenvalue decomposition. For example, if a matrix is not square, the eigendecomposition is not defined

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\top}$$

- U and V are both defined to be orthogonal matrices, while D is diagonal

- The elements along the diagonal of D are known as **the singular values of the matrix A**. The columns of U are known as the **left-singular** vectors. The columns of V are known as as the **right-singular** vectors.

- Perhaps the most useful feature of the SVD is that we can use it to partially **generalize matrix inversion** to **non-square** matrices

Statistics and probability:

- Random variable: a variable that can take on different values randomly.

Statistics and probability:

- Random variable: a variable that can take on different values randomly.

- A random variable is just a description of the states that are possible; it must be coupled with a probability distribution that specifies how likely each of these states are.

- A probability distribution is a description of how likely a random variable or set of random variables is to take on each of its possible states.

Statistics and probability:

- Random variable: a variable that can take on different values randomly.
- A random variable is just a description of the states that are possible; it must be coupled with a probability distribution that specifies how likely each of these states are.
- A probability distribution is a description of how likely a random variable or set of random variables is to take on each of its possible states.
- Probability Mass Function (PMF) : A probability distribution over discrete variables, a function maps from a state of a random variable to the probability of that random variable taking on that state

$$\sum_{x \in \mathrm{x}} P(x) = 1$$

- E.g PMF of uniform distribution = $P(\mathrm{x} = x_i) = \frac{1}{k}$ where k is the number of the states

Statistics and probability:

- Joint probability distribution. P(x,y) denotes the probability that x = x and y = y simultaneously.

- probability density function (PDF) describe probability distribution of continuous random variable

$$p(x) \leq 1 \qquad \int p(x)dx = 1.$$

- PDF **does not** give the probability of a specific state directly, instead the probability of being within interval [a, b] by the integral $\int_{[a,b]} p(x)dx.$

- Example Uniform distribution ***U***(x;a,b)

Statistics and probability:

- Joint probability distribution. P(x,y) denotes the probability that x = x and y = y simultaneously.

- probability density function (PDF) describe probability distribution of continuous random variable

$$p(x) \leq 1 \qquad \int p(x)dx = 1.$$

• PDF **does not** give the probability of a specific state directly, instead the probability of being within interval [a, b] by the integral $\int_{[a,b]} p(x)dx.$

- Example Uniform distribution ***U***(x;a,b)

- Marc course ☺ for details on common probability distributions.

Statistics and probability:

Summary statistics:

- **Expectation**: expected value of some function f(x) with respect to a probability distribution P (x) is the average or mean value that f takes on when x is drawn from P .

$$\mathbb{E}_{\mathrm{x}\sim p}[f(x)] = \int p(x)f(x)dx.$$

Statistics and probability:

Summary statistics:

- **Expectation**: expected value of some function f(x) with respect to a probability distribution P (x) is the average or mean value that f takes on when x is drawn from P .

$$\mathbb{E}_{\mathrm{x}\sim p}[f(x)] = \int p(x)f(x)dx.$$

- **Variance** : gives a measure of how much the values of a function of a random variable x vary as we sample different values of x from its probability distribution:

$$\mathrm{Var}(f(x)) = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right].$$

- Square root of the variance is known as the standard deviation.

Statistics and probability:

Summary statistics:

- **Covariance**: gives some sense of how much two values are linearly related to each other, as well as the scale of these variables:

$$\mathrm{Cov}(f(x), g(y)) = \mathbb{E}\left[(f(x) - \mathbb{E}\left[f(x)\right])\left(g(y) - \mathbb{E}\left[g(y)\right]\right)\right].$$

- **Correlation**: normalize the contribution of each variable in order to measure only how much the variables are related, rather than also being affected by the scale of the separate variables**.**

Statistics and probability:

Summary statistics:

- **Covariance**: gives some sense of how much two values are linearly related to each other, as well as the scale of these variables:

$$\mathrm{Cov}(f(x), g(y)) = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])(g(y) - \mathbb{E}[g(y)])\right].$$

- **Correlation**: normalize the contribution of each variable in order to measure only how much the variables are related, rather than also being affected by the scale of the separate variables**.**

- **Dependence**: It is possible for two variables to be dependent but have zero covariance:

  - For two variables to have zero covariance, there must be no linear dependence between them

  - Independence is a stronger requirement than zero covariance, because independence also excludes nonlinear relationships.

  - It is possible for two variables to be dependent but have zero covariance

Statistics and probability:

- **The central limit theorem**: shows that the sum of many independent random variables is approximately normally distributed

- This means that in practice, many complicated systems can be modeled successfully as normally distributed noise, even if the system can be decomposed into parts with more structured behavior.

Statistics and probability:

- Why Statistics is important in machine learning?

1. Statistical machine learning.
2. Summarize datasets (observations) into information.
3. Formulate, and reason machine learning problem as task with statistical objective e.g generative model.
4. Quantify, and describe your results and performance metrics.
5. Borrow statistical techniques and methods to solve sub tasks.

References

- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

- Mathematics for Machine Learning, A. Aldo Faisal, Cheng Soon Ong, and Marc Peter Deisenroth