CMSC 422 Introduction to Machine Learning

Lecture 2 Reviews

Furong Huang / furongh@umd.edu

UNIVERSITY OF MARYLAND

Reviews
- Linear Algebra

Collection of vectors, Subspaces

A set of T-vectors, $V = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_T\}$

- **Linear combination:** $\sum_{k=1}^{T} \alpha_k \mathbf{a}_k, \alpha_k \in \mathbb{R}$
- **Linearly independent:** No vector in V can be written as linear combination of others
- **Span:** $\text{Span}(V) = \{\mathbf{x} \,|\, \mathbf{x} = \sum_k \alpha_k \mathbf{a}_k, \alpha_k \in \mathbb{R}\}$

Collection of vectors, Subspaces

Definition (Subspace)

A collection of vectors $V \subset \mathbb{R}^N$ is a subspace iff it is closed under linear combinations

$$\mathbf{a}, \mathbf{b} \in V \implies \alpha\mathbf{a} + \beta\mathbf{b} \in V, \alpha, \beta \in \mathbb{R}$$

- **Basis of a subspace:** A linearly independent *spanning* set
- **Dimensionality of a subspace:** #elements in a basis

Matrix

- $A \in \mathbb{R}^{M \times N}$: A matrix of dimension $M \times N$
- $A = [a_{ij}] = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N], \mathbf{a}_i \in \mathbb{R}^M$
- $\mathrm{rank}(A) =$ largest number of linearly *independent* columns
- $\mathrm{rank}(A) = \mathrm{rank}(A^T) \leq \min(M, N)$
- A is *full-rank* if $\mathrm{rank}(A) = \min(M, N)$.

Matrices are *representations of linear operators.*

$A : \mathbb{R}^N \to \mathbb{R}^M$
$\mathbf{x} \in \mathbb{R}^N \mapsto A\mathbf{x} \in \mathbb{R}^M$

Eigenvectors and Eigenvalues

- Let A be a $N \times N$ square matrix
- $\mathbf{x}$ is an eigenvector and λ is an eigenvalue of A is

$$A\mathbf{x} = \lambda\mathbf{x}$$

- **Intuition:** eigenvectors are vectors in $\mathbb{R}^N$ whose direction is preserved under action of A; however, length may change
- Eigen-decomposition: $A = UDU^{-1}$

Spectral Theorem

Theorem

If $A = A^H$, then

- *The matrix is "symmetric"*
- *all eigenvalues are real*
- *eigenvectors with different eigenvalues are perpendicular*
- *there exists a complete orthogonal basis of eigenvectors.*

Singular value decomposition (SVD)

Definition (SVD)

Any matrix $A \in \mathbb{R}^{M \times N}$ can be written as

$$A = U \Sigma V^T,$$

where $U \in \mathbb{R}^{M \times M}$ and $V \in \mathbb{R}^{N \times N}$ are unitary and $\Sigma \in \mathbb{R}^{M \times N}$ is diagonal.

- Diagonal entries of $\Sigma = \{\sigma_i\}$ are called the singular values; they are positive dand real. Typically, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$
- Singular values are the eigenvalues of $\sqrt{A^T A}$ and $\sqrt{A A^T}$.
- If $A = A^T$, singular values are same as the eigenvalues
- The ratio of the largest to smallest singular value is the so-called *condition number* of A

Reviews
- Probability Theory

Probability theory – study of uncertainty

- Axioms of Probability
  - Sample space Ω: the set of all the outcomes of a random experiment.
  - Event space $\boldsymbol{\mathcal{F}}$: A set whose elements $A \in \boldsymbol{\mathcal{F}}$ (called events) are subsets of Ω
  - Probability measure: A function $P: \boldsymbol{\mathcal{F}} \rightarrow \mathbb{R}$ that satisfied the following properties
    - $P(A) \geq 0$, for all $A \in \boldsymbol{\mathcal{F}}$
    - $P(\Omega) = 1$
    - If $A_1, A_2, \ldots$ are disjoint events, then $P(\cup_i A_i) = \sum_i P(A_i)$

Random Variables

- A random variable X is a function $X: \Omega \to \mathbb{R}$
- Cumulative distribution functions
  - CDF (cumulative distribution function) $F_X: \mathbb{R} \to [0,1]$ which specifies a probability measure as $F_X(x) \coloneqq P(X \leq x)$
  - PMF (probability mass function) for discrete random variable $p_X: \Omega \to \mathbb{R}$ such that $p_X(x) \coloneqq P(X = x)$
  - PDF (probability density function) for continuous random variables such that the CDF is differentiable everywhere,. PDF is the derivative of CDF $f_X(x) \coloneqq \frac{dF_X(x)}{dx}$
- Expectation $\mathbb{E}[g(X)] \coloneqq \sum_x g(x)p_X(x)$ or $\mathbb{E}[g(x)] \coloneqq \int_{-\infty}^{\infty} g(x)f_X(x)dx$
- Variance $Var[X] \coloneqq \mathbb{E}[(x - \mathbb{E}[x])^2]$

Common Random Variables

- $X \sim Bernoulli(p), p(x) = \begin{cases} p & if\ x = 1 \\ 1 - p & if\ x = 0 \end{cases}$
- $X \sim Binomial(n, p), p(x) = \binom{n}{x} \mathrm{p}^{\mathrm{x}}(1 - \mathrm{p})^{n-x}$, x is the number of heads in n independent flips of a coin with heads probability p
- $X \sim Geometric(p), p(x) = p(1-p)^{x-1}$, x is the number of flips of a coin with heads probability p until the first heads
- $X \sim Poisson(\lambda),\ p(x) = e^{-\lambda}\frac{\lambda^x}{x!}$ a probability distribution over the nonnegative integers used for modeling the frequency of rare events.

Common Random Variables

- $X \sim Uniform(a, b), f(x) = \begin{cases} \frac{1}{b-a} & if\ a \leq x \leq b \\ 0 & o.w. \end{cases}$
- $X \sim Exponential(\lambda),\ f(x) = \begin{cases} \lambda e^{-\lambda x} & if\ x \geq 0 \\ 0 & o.w. \end{cases}$
- $X \sim Normal(\mu, \sigma^2), f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$

Multiple Random Variables $X_1, X_2, \ldots, X_n$

- Joint distribution function

$$F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$$

- Joint probability density function

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots x_n) = \frac{\partial^n F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots x_n)}{\partial x_1 \ldots \partial x_n}$$

- Marginal probability density function

$$f_{X_1}(X_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots x_n) dx_2 \ldots dx_n$$

- Conditional probability density function

$$f_{X_1|X_2,\ldots,X_n}(x_1|x_2, \ldots x_n) = \frac{f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots x_n)}{f_{X_2,\ldots,X_n}(x_1, x_2, \ldots x_n)}$$

Chain rule

$$
\begin{aligned}
f(x_1, x_2, \ldots, x_n) &= f(x_n | x_1, x_2 \ldots, x_{n-1}) f(x_1, x_2 \ldots, x_{n-1}) \\
&= f(x_n | x_1, x_2 \ldots, x_{n-1}) f(x_{n-1} | x_1, x_2 \ldots, x_{n-2}) f(x_1, x_2 \ldots, x_{n-2}) \\
&= \ldots = f(x_1) \prod_{i=2}^{n} f(x_i | x_1, \ldots, x_{i-1}).
\end{aligned}
$$

Independence

$$
f(x_1, \ldots, x_n) = f(x_1) f(x_2) \cdots f(x_n)
$$

- Independent random variables arise often in machine learning algorithms where we assume that the training examples belonging to the training set represent independent samples from some unknown probability distribution.
- To make the significance of independence clear, consider a "bad" training set in which we first sample a single training example $(x^{(1)}, y^{(1)})$ from some unknown distribution, and then add m − 1 copies of the exact same training example to the training set. In this case, the examples are not independent

$$P((x^{(1)}, y^{(1)}), \ldots . (x^{(m)}, y^{(m)})) \neq \prod_{i=1}^{m} P(x^{(i)}, y^{(i)})$$

in practice, non-independence of samples does come up often, and it has the effect of reducing the "effective size" of the training set.

Random Vectors

- $X = [X_1, X_2, \ldots, X_n]^\top : \Omega \to \mathbb{R}^n$, an alternative notation for dealing with n random variables, so the notions of joint PDF and CDF will apply to random vectors as well.
- A special case: n independent Bernoulli variables $X_i \sim Bernoulli(p_i)$
  - $\mathbb{E}[X] = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{bmatrix} = \begin{bmatrix} p_1 \\ \vdots \\ p_n \end{bmatrix} = \sum_i p_i \, \overrightarrow{e_i}$
- Multivariate Gaussian with mean $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{S}^n_{++}$(space of symmetric positive definite $n \times n$ matrices)

$$f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right).$$

Gaussian random variables are extremely useful in machine learning and statis tics for two main reasons.

- First, they are extremely common when modeling “noise” in statistical algorithms. Quite often, noise can be the accumulation of many small independent random perturbations affecting the measurement process; by the Central Limit Theorem, summations of independent random variables will tend to “look Gaussian.”
- Second, Gaussian random variables are convenient for many analytical manipulations, because many of the integrals involving Gaussian distributions that arise in practice have simple closed form solutions.

Reviews
- (Multidimensional) Calculus

Calculus

Classically, calculus is the study of the relationship between variables and their rates of change. However, this is not what we use calculus for.

We use differential calculus as a method for finding extrema of functions; we use integral calculus as a method for probabilistic modeling.

Differential Calculus

- The derivative of a function at a point is the slope of the function at that point. The derivative of f with respect to x, denoted as $\partial f / \partial x$, is

$$f'(x_0) = \frac{\partial f}{\partial x}(x_0) = \lim_{n \to 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

- Scalar multiplication: $\partial_x[af(x)] = a[\partial_x f(x)]$
- Polynomials: $\partial_x[x^k] = kx^{k-1}$
- Function addition: $\partial_x[f(x) + g(x)] = [\partial_x f(x)] + [\partial_x g(x)]$
- Function multiplication: $\partial_x[f(x)g(x)] = f(x)[\partial_x g(x)] + [\partial_x f(x)]g(x)$
- Function division: $\partial_x\left[\frac{f(x)}{g(x)}\right] = \frac{[\partial_x f(x)]g(x) - f(x)[\partial_x g(x)]}{[g(x)]^2}$
- Function composition: $\partial_x[f(g(x))] = [\partial_x g(x)][\partial_x f](g(x))$
- Exponentiation: $\partial_x[e^x] = e^x \quad \text{and} \quad \partial_x[a^x] = \log(a)e^x$
- Logarithms: $\partial_x[\log x] = \frac{1}{x}$

Multidimensional Equivalents of Derivatives

- Suppose we have vectors $\boldsymbol{w}$ and $\boldsymbol{x}$ and a function $f(\boldsymbol{w}, \boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$. We might want to differentiate f with respect to w so that we can minimize f --- differentiate with respect to vectors.
- $\boldsymbol{w} = [w_1, w_2, \ldots, w_D]^\top \in \mathbb{R}^D$ and $f(\boldsymbol{w}) \in \mathbb{R}$, we compute the partial derivatives of f with respect to each component of $\boldsymbol{w}$

  $\frac{\partial f}{\partial w_1}, \frac{\partial f}{\partial w_2}, \frac{\partial f}{\partial w_3}, \ldots, \frac{\partial f}{\partial w_D}$, each of these are just univariate derivatives
- Package them together into a vector, we get the **gradient** of f with respect to the vector $\boldsymbol{w}$, denoted as $\nabla_{\boldsymbol{w}} f$

$$\nabla_{\boldsymbol{w}} f = \begin{bmatrix} \frac{\partial f}{\partial w_1} \\ \frac{\partial f}{\partial w_2} \\ \frac{\partial f}{\partial w_3} \\ \vdots \\ \frac{\partial f}{\partial w_D} \end{bmatrix}$$

The gradient of a function is its velocity in each available dimension.

Multidimensional Equivalents of Second Derivatives

- Hessian matrix

$$\mathbf{H}(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial w_1^2} & \frac{\partial^2 f}{\partial w_1 \partial w_2} & \cdots & \frac{\partial^2 f}{\partial w_1 \partial w_D} \\ \frac{\partial^2 f}{\partial w_1 \partial w_2} & \frac{\partial^2 f}{\partial w_2^2} & \cdots & \frac{\partial^2 f}{\partial w_2 \partial w_D} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 f}{\partial w_1 \partial w_D} & \frac{\partial^2 f}{\partial w_2 \partial w_D} & \cdots & \frac{\partial^2 f}{\partial w_D^2} \end{bmatrix}$$

$$\Big(\mathbf{H}(f)\Big)_{i,j} = \frac{\partial^2 f}{\partial w_i \partial w_j}.$$

The second derivative of a function is the rate of change of its velocity, aka its acceleration.

The Hessian of a function is the rate at which different dimension accelerate together.

Matrix Cookbook

- https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

Integral Calculus

- “Opposite” of a derivative, commonly used for computing areas under a curve
- Continuous analogues of simple sums
- The interpretation of an integral as a sum:
  - Suppose we were to discretize the range [a, b] into R many units of width (a − b)/R.
  - Then, we could approximate the area under the curve by a sum over these units, evaluating f(x) at each position (to get the height of a rectangle there) and multiplying by the width (a − b)/R.
  - Summing these rectangles will approximate the area. As we let $R \to \infty$, we’ll get a better and better approximation.
  - However, as $R \to \infty$, the width of each rectangle will approach 0. We name this width “dx,” and thus the integral notation mimics almost exactly the “rectangular sum” notation (we have width of dx times height of f(x), summed over the range).

- Check out course webpage, Canvas, Piazza
- Submit HW01
  - due Tuesday 10:30am
- Do the readings!

UNIVERSITY OF MARYLAND

Furong Huang
4124 IRB, College Park, MD 20740
301.405.8010 / furongh@umd.edu