UCLA Samueli Computer Science

CS145 Discussion: Week 1

Shichang, Yewen, Zongyue

Friday, 09/24/2021

Roadmap

- Course logistics
- Math review
  - Probability
  - Linear algebra
  - Optimization
  - Matrix calculus
- Python and Google Colab set up

Roadmap

- **Course logistics**
- Math review
  - Probability
  - Linear algebra
  - Optimization
  - Matrix calculus
- Python and Google Colab set up

Course logistics

- Course homepage:
  - https://github.com/yichousun/Fall2021_CS145__IntroDM
  - Please find all the relevant course information there, e.g. schedule, slides, and etc.
- Piazza:
  - piazza.com/ucla/fall2021/cs145
  - Please ask your question on Piazza **before** email the professor or any TAs, so others will also benefit from your question.
- Important dates
  - Midterm: 11/4 (Thursday)
  - Final exam: 12/9 (Thursday)
  - First homework out and project details: next week in discussion

Course logistics

- **Office hours**
  - **Yizhou Sun** (yzsun@cs.ucla.edu) Monday 2-3 and Tuesday 4:15-5:00 @ zoom
  - **Shichang Zhang** (shichang@cs.ucla.edu), office hours: Friday 10am-12pm @ BH 3551 Conference Room (May change to the TA office BH 3256 once it is open)
  - **Yewen Wang** (wyw10804@gmail.com), office hours: Wednesday 9-10am @ Boelter Hall 3551 Conference Room, 10-11am @ zoom
  - **Zongyue Qin** (qinzongyue@cs.ucla.edu), office hours: Monday 9-11am @ BH 3551 (row M)

Roadmap

- Course logistics
- **Math review**
  - Probability
  - Linear algebra
  - Optimization
  - Matrix calculus
- Python and Google Colab set up

Math Review

- Slides reference
  - Jeff Howbert, https://courses.washington.edu/css490/2012.Winter/lecture_slides/02_math_essentials.pdf
  - Xinkun Nie, http://cs229.stanford.edu/notes2020fall/notes2020fall/TA-slides1.pdf
  - Hristo Paskov, http://snap.stanford.edu/class/cs246-2014/slides/LinAlgSession.pdf

Probability

Random variables

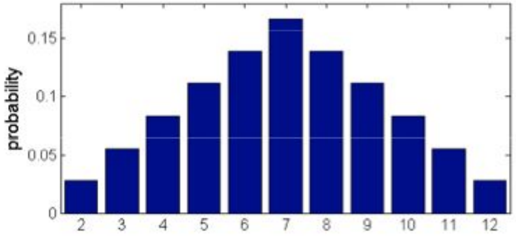
- A random variable X is a function that associates a number x with each outcome O of a process
  - Common notation: $X(O) = x$, or just $X = x$
- Example: X = number of heads in three flips of a coin
  - Possible values of X are 0, 1, 2, 3
  - $p(X = 0) = p(X = 3) = 1 / 8 \qquad p(X = 1) = p(X = 2) = 3 / 8$
  - Size of space (number of “outcomes”) reduced from 8 to 4
- Example: X = average height of five randomly chosen American men
  - Size of space unchanged (X can range from 2 feet to 8 feet), but pdf of X different than for single man

Probability

Probability distributions

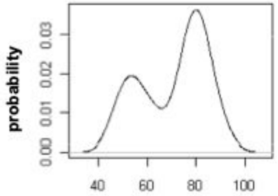
- Discrete: *probability mass function* (pmf)

  example:
  sum of two
  fair dice



- Continuous: *probability density function* (pdf)

  example:
  waiting time between
  eruptions of Old Faithful
  (minutes)



Probability

Multivariate probability distributions

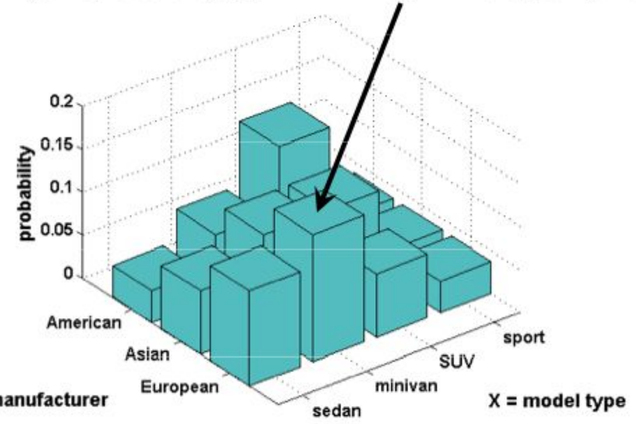
- Scenario
  - Several random processes occur (doesn't matter whether in parallel or in sequence)
  - Want to know probabilities for each possible combination of outcomes
- Can describe as *joint probability* of several random variables
  - Example: two processes whose outcomes are represented by random variables X and Y. Probability that process X has outcome x and process Y has outcome y is denoted as:

$$p(X = x, Y = y)$$

Probability

Example of multivariate distribution

joint probability: $p(X = \text{minivan}, Y = \text{European}) = 0.1481$



Probability

Multivariate probability distributions

- *Marginal* probability
  - Probability distribution of a single variable in a joint distribution
  - Example: two random variables X and Y:

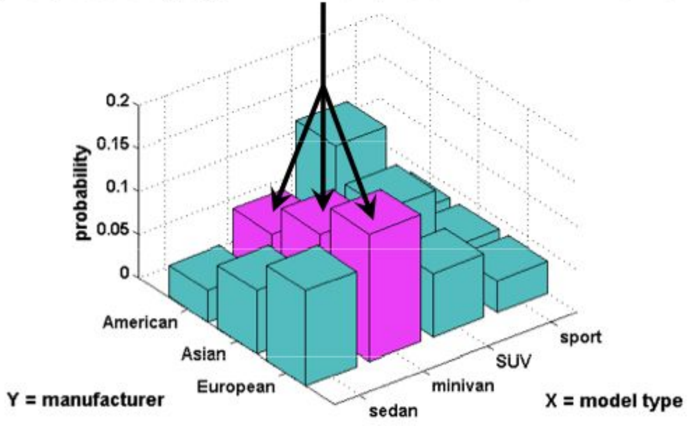
    $p(X = x) = \sum_{b = \text{all values of } Y} p(X = x, Y = b)$
- *Conditional* probability
  - Probability distribution of one variable *given* that another variable takes a certain value
  - Example: two random variables X and Y:

    $p(X = x \mid Y = y) = p(X = x, Y = y) / p(Y = y)$

Probability

Example of marginal probability

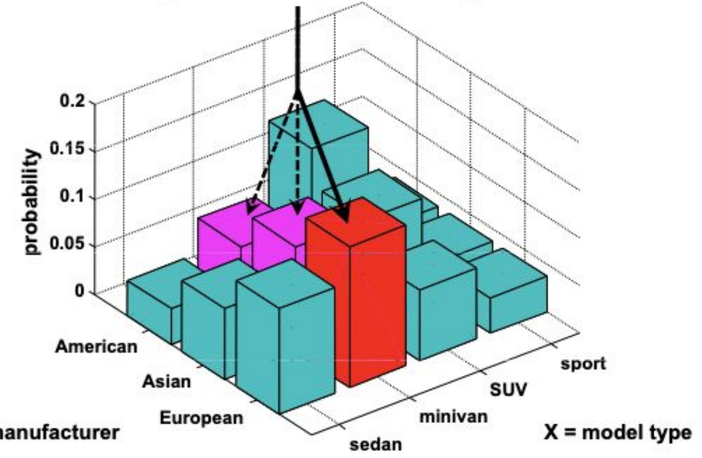
marginal probability: $p(X = \text{minivan}) = 0.0741 + 0.1111 + 0.1481 = 0.3333$



Probability

Example of conditional probability

conditional probability: $p(\ Y = \text{European} \mid X = \text{minivan}\ ) =$
$0.1481\ /\ (\ 0.0741 + 0.1111 + 0.1481\ ) = 0.4433$



Probability

Expected value

Given:

- A discrete random variable X, with possible values $x = x_1, x_2, \ldots x_n$
- Probabilities $p(X = x_i)$ that X takes on the various values of x_i
- A function $y_i = f(x_i)$ defined on X

The *expected value* of f is the probability-weighted "average" value of $f(x_i)$:

$$E(f) = \sum_i p(x_i) \cdot f(x_i)$$

Probability

Common forms of expected value (1)

- Mean (μ)

  $f(x_i) = x_i \Rightarrow \quad \mu = E(f) = \sum_i p(x_i) \cdot x_i$

  - Average value of $X = x_i$, taking into account probability of the various x_i
  - Most common measure of “center” of a distribution

Probability

Common forms of expected value (2)

- Variance (σ^2)

  $f(x_i) = (x_i - \mu) \quad \Rightarrow \qquad \sigma^2 = \sum_i p(x_i) \cdot (x_i - \mu)^2$

  - Average value of squared deviation of $X = x_i$ from mean μ, taking into account probability of the various x_i
  - Most common measure of “spread” of a distribution
  - σ is the *standard deviation*

Probability

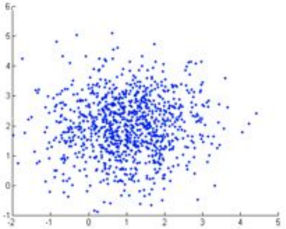
Common forms of expected value (3)

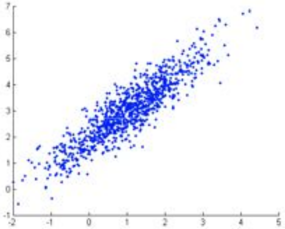
- Covariance

  $f(x_i) = (x_i - \mu_x),\quad g(y_i) = (y_i - \mu_y) \quad \Rightarrow$
  $\mathrm{cov}(x, y) = \sum_i p(x_i, y_i) \cdot (x_i - \mu_x) \cdot (y_i - \mu_y)$

  – Measures tendency for *x* and *y* to deviate from their means in same (or opposite) directions at same time

no covariance





high (positive) covariance

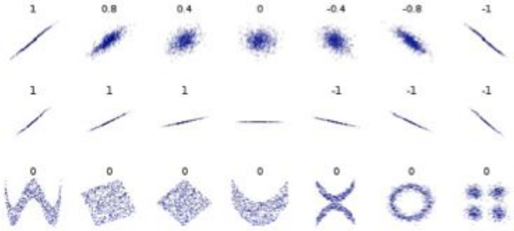
Probability

Correlation

- Pearson's correlation coefficient is covariance normalized by the standard deviations of the two variables

$$\mathrm{corr}(x, y) = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}$$

  - Always lies in range -1 to 1
  - Only reflects *linear dependence* between variables



Linear dependence with noise

Linear dependence without noise

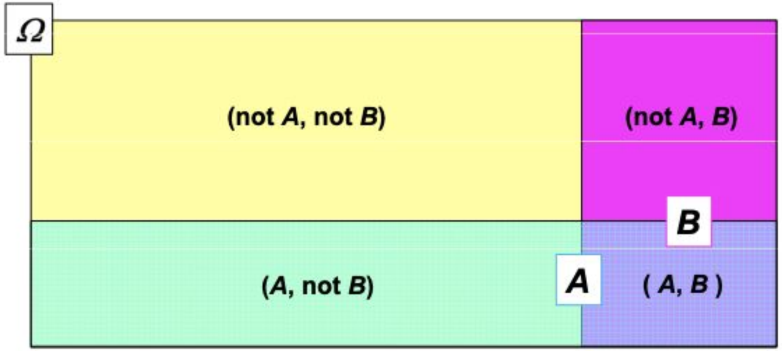
Various nonlinear dependencies

Probability

Independence

Given: events A and B, which can co-occur (or not)

$$p(A \mid B) = p(A) \quad \text{or} \quad p(A, B) = p(A) \cdot p(B)$$



areas represent relative probabilities

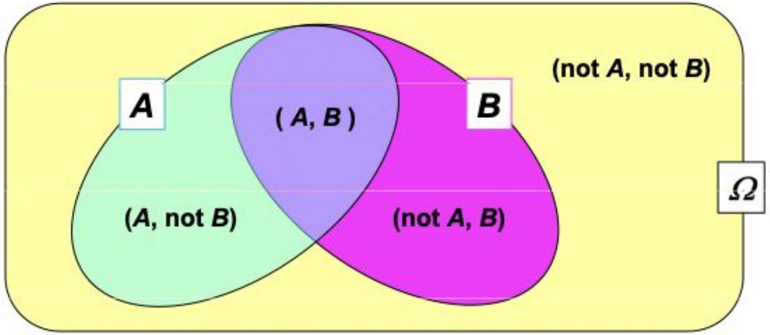
Probability

Bayes rule

A way to find conditional probabilities for one variable when conditional probabilities for another variable are known.

$$p(B \mid A) = p(A \mid B) \cdot p(B) / p(A)$$

where $p(A) = p(A, B) + p(A, \text{not } B)$

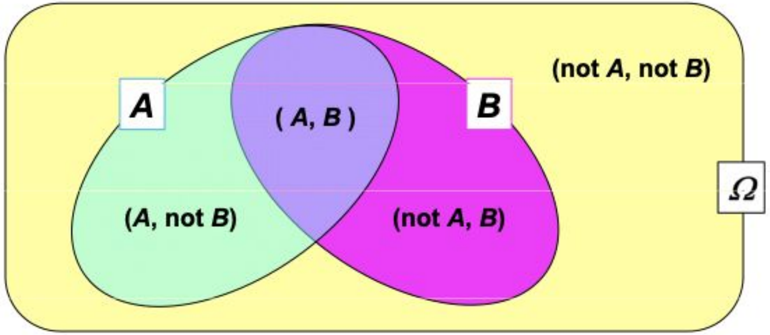


Probability

Bayes rule

posterior probability $\propto$ likelihood $\times$ prior probability

$$p(B \mid A) = p(A \mid B) \cdot p(B) \,/\, p(A)$$



Roadmap

- Course logistics
- **Math review**
  - Probability
  - Linear algebra
  - Optimization
  - Matrix calculus
- Python and Google Colab set up

Vectors and Matrices

- Vector $x \in \mathbb{R}^d$

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

- May also write

$$x = [x_1 \quad x_2 \quad \dots \quad x_d]^T$$

Vectors and Matrices

- Matrix $M \in \mathbb{R}^{m \times n}$

$$M = \begin{bmatrix} M_{11} & \cdots & M_{1n} \\ \vdots & \ddots & \vdots \\ M_{m1} & \cdots & M_{mn} \end{bmatrix}$$

- Written in terms of rows or columns

$$M = \begin{bmatrix} \boldsymbol{r}_1^T \\ \vdots \\ \boldsymbol{r}_m^T \end{bmatrix} = [\boldsymbol{c}_1 \quad \dots \quad \boldsymbol{c}_n]$$

$$\boldsymbol{r}_i = [M_{i1} \quad \dots \quad M_{in}]^T \quad \boldsymbol{c}_i = [M_{1i} \quad \dots \quad M_{mi}]^T$$

Linear algebra

Multiplication

- Vector-vector: $x, y \in \mathbb{R}^d \to \mathbb{R}$

$$x^T y = \sum_{i=1}^{d} x_i y_i$$

- Matrix-vector: $x \in \mathbb{R}^n, M \in \mathbb{R}^{m \times n} \to \mathbb{R}^m$

$$Mx = \begin{bmatrix} \boldsymbol{r}_1^T \\ \vdots \\ \boldsymbol{r}_m^T \end{bmatrix} x = \begin{bmatrix} \boldsymbol{r}_1^T x \\ \vdots \\ \boldsymbol{r}_m^T x \end{bmatrix}$$

By columns $\begin{bmatrix} 2 & 3 \\ 2 & 4 \\ 3 & 7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1 \begin{bmatrix} 2 \\ 2 \\ 3 \end{bmatrix} + x_2 \begin{bmatrix} 3 \\ 4 \\ 7 \end{bmatrix} =$ combination of the columns $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$

Matrix Column Space

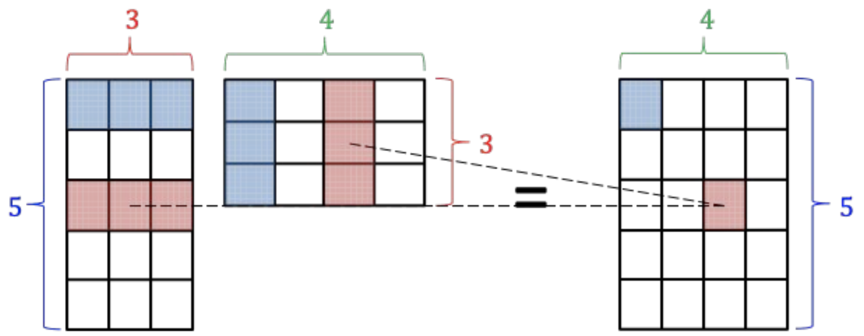
$\boxed{\textbf{$Ax$ is a linear combination of the columns of A. This is fundamental.}}$

[1] More about Matrix Analysis

Linear algebra

Multiplication

- Matrix-matrix: $A \in \mathbb{R}^{m \times k}, B \in \mathbb{R}^{k \times n} \rightarrow \mathbb{R}^{m \times n}$



$$\boldsymbol{A} = \begin{bmatrix} 1 & 3 & 8 \\ 1 & 2 & 6 \\ 0 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 2 \end{bmatrix}$$

Linear algebra

Multiplication Properties

- Associative

$$(AB)C = A(BC)$$

- Distributive

$$A(B + C) = AB + BC$$

- NOT commutative

$$AB \neq BA$$

  – Dimensions may not even be conformable

Useful Matrices

- Identity matrix $I \in \mathbb{R}^{m \times m}$
  - $AI = A, IA = A$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad I_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

- Diagonal matrix $A \in \mathbb{R}^{m \times m}$

$$A = \mathrm{diag}(a_1, \dots, a_m) = \begin{bmatrix} a_1 & \cdots & 0 \\ \vdots & a_i & \vdots \\ 0 & \cdots & a_m \end{bmatrix}$$

Linear algebra

Useful Matrices

- Symmetric $A \in \mathbb{R}^{m \times m}$: $A = A^T$
- Orthogonal $U \in \mathbb{R}^{m \times m}$:
$$U^T U = U U^T = I$$
  – Columns/ rows are orthonormal
- Positive semidefinite $A \in \mathbb{R}^{m \times m}$:
$$x^T A x \geq 0 \quad \text{for all } x \in \mathbb{R}^m$$
  – Equivalently, there exists $L \in \mathbb{R}^{m \times m}$
$$A = LL^T$$

Properties of real symmetric matrix

1. Eigenvalues are real.
2. Eigenvectors of different eigenvalues are orthogonal

Norms

- Quantify "size" of a vector
- Given $x \in \mathbb{R}^n$, a norm satisfies
  1. $\|cx\| = |c|\|x\|$
  2. $\|x\| = 0 \Leftrightarrow x = 0$
  3. $\|x + y\| \leq \|x\| + \|y\|$
- Common norms:
  1. Euclidean L_2-norm: $\|x\|_2 = \sqrt{x_1^2 + \cdots + x_n^2}$
  2. L_1-norm: $\|x\|_1 = |x_1| + \cdots + |x_n|$
  3. L_∞-norm: $\|x\|_\infty = \max_i |x_i|$

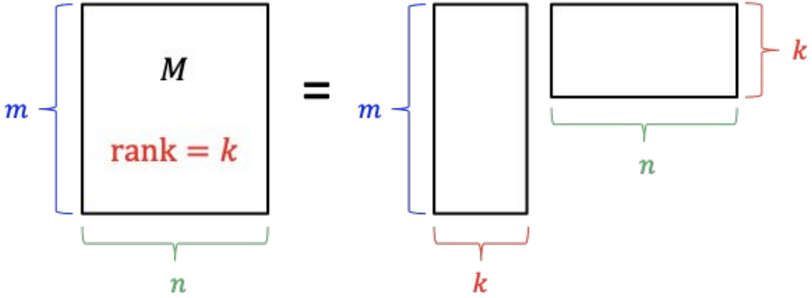
UCLA Engineer Change.

Matrix Rank

- the **rank** of a matrix A is the dimension of the vector space generated (or spanned) by its columns.
- This corresponds to the maximal number of linearly independent columns of A.
- This, in turn, is identical to the dimension of the vector space spanned by its rows

Matrix Rank

- $\text{rank}(M)$ gives dimensionality of row <u>and</u> column spaces
- If $M \in \mathbb{R}^{m \times n}$ has rank k, can decompose into product of $m \times k$ and $k \times n$ matrices



Properties of Rank

- For $A, B \in \mathbb{R}^{m \times n}$
  1. $\text{rank}(A) \leq \min(m, n)$
  2. $\text{rank}(A) = \text{rank}(A^T)$
  3. $\text{rank}(AB) \leq \min\big(\text{rank}(A), \text{rank}(B)\big)$
  4. $\text{rank}(A + B) \leq \text{rank}(A) + \text{rank}(B)$
- A has *full rank* if $\text{rank}(A) = \min(m, n)$
- If $m > \text{rank}(A)$ rows not linearly independent
  – Same for columns if $n > \text{rank}(A)$

Matrix Inverse

- $M \in \mathbb{R}^{m \times m}$ is invertible iff $\mathrm{rank}(M) = m$
- Inverse is unique and satisfies
  1. $M^{-1}M = MM^{-1} = I$
  2. $(M^{-1})^{-1} = M$
  3. $(M^T)^{-1} = (M^{-1})^T$
  4. If A is invertible then MA is invertible and $(MA)^{-1} = A^{-1}M^{-1}$

Linear algebra

Characterizations of Eigenvalues

- Traditional formulation

$$Mx = \lambda x$$

  - Leads to characteristic polynomial

$$\det(M - \lambda I) = 0$$

Linear algebra

Eigenvalue Properties

- For $M \in \mathbb{R}^{m \times m}$ with eigenvalues λ_i
  1. $\mathrm{tr}(M) = \sum_{i=1}^{m} \lambda_i$
  2. $\det(M) = \lambda_1 \lambda_2 \ldots \lambda_m$
  3. $\mathrm{rank}(M) = \#\lambda_i \neq 0$

Roadmap

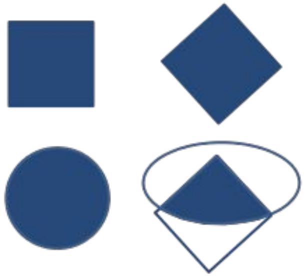
- Course logistics
- **Math review**
  - Probability
  - Linear algebra
  - Optimization
  - Matrix calculus
- Python and Google Colab set up

Convex Sets

- A set C is convex if $\forall x, y \in C$ and $\forall \alpha \in [0,1]$

$$\alpha x + (1-\alpha)y \in C$$

  - Line segment between points in C also lies in C
- Ex
  - Intersection of halfspaces
  - L_p balls
  - Intersection of convex sets



Convex Functions

- A real-valued function f is convex if $\mathrm{dom} f$ *is* convex and $\forall x, y \in \mathrm{dom} f$ and $\forall \alpha \in [0,1]$

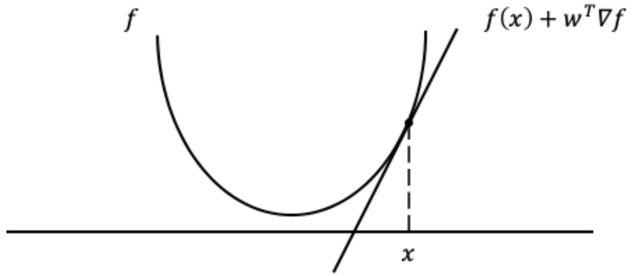
$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$$

  – Graph of f upper bounded by line segment between points on graph

Gradients

- Differentiable convex f with $\mathrm{dom} f = \mathbb{R}^d$
- Gradient ∇f at x gives linear approximation

$$\nabla f = \left[\frac{\delta f}{\delta x_1} \quad \ldots \quad \frac{\delta f}{\delta x_d}\right]^T$$

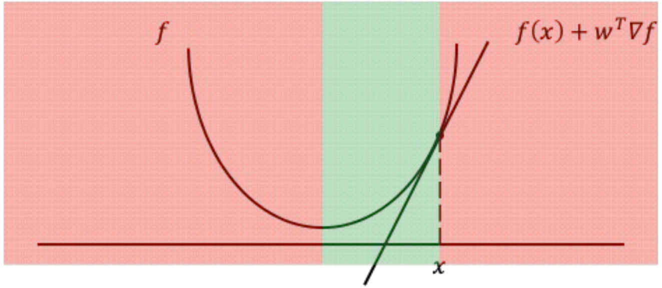


Optimization

Gradients

- Differentiable convex f with $\mathrm{dom} f = \mathbb{R}^d$
- Gradient ∇f at x gives linear approximation

$$\nabla f = \left[\frac{\delta f}{\delta x_1} \quad \dots \quad \frac{\delta f}{\delta x_d}\right]^T$$



Optimization

Gradient Descent

- To minimize f move down gradient
  - But not too far!
  - Optimum when $\nabla f = 0$
- Given f, learning rate α, starting point x_0

  $x = x_0$

  Do until $\nabla f = 0$

  $x = x - \alpha \nabla f$

Optimization

Stochastic Gradient Descent

- Given $f(\theta) = \sum_{i=1}^{n} L(\theta; \boldsymbol{x}_i)$, learning rate α, starting point θ_0

  $\theta = \theta_0$

  Do until $f(\theta)$ nearly optimal

  For $i = 1$ to n in random order

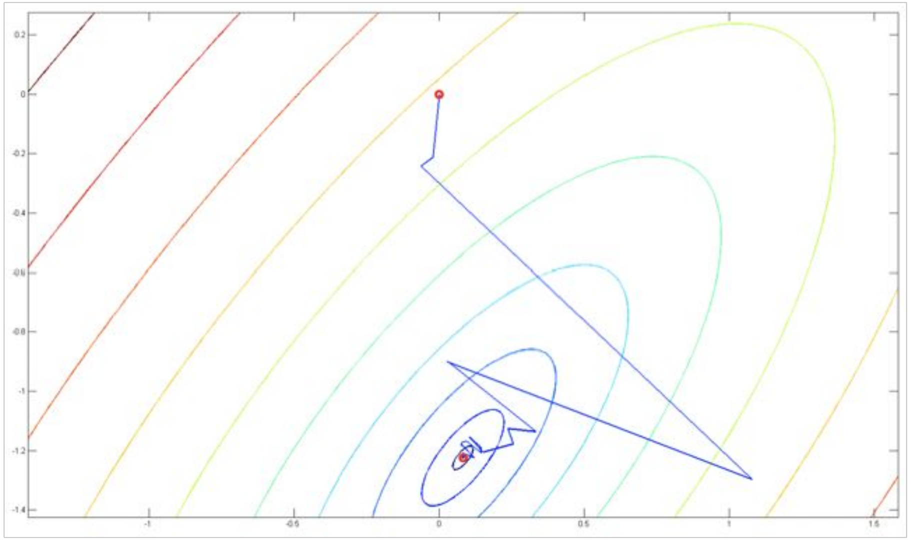
  $$\theta = \theta - \alpha \nabla L(\theta; \boldsymbol{x}_i)$$

- Finds nearly optimal θ

What if have 3 million datapoints? SGD used the cost gradient of **1 example** at each iteration, instead of using the sum of the cost gradient of **ALL** examples

Optimization

Minimize $\sum_{i=1}^{n}(y_i - \theta^T \boldsymbol{x}_i)^2$



Roadmap

- Course logistics
- **Math review**
  - Probability
  - Linear algebra
  - Optimization
  - Matrix calculus
- Python and Google Colab set up

Matrix Calculus

A helpful link:

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

The Gradient

Suppose that $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is a function that takes as input a matrix A of size $m \times n$ and returns a real value. Then the ***gradient*** of f (with respect to $A \in \mathbb{R}^{m\times n}$) is the matrix of partial derivatives, defined as:

$$\nabla_A f(A) \in \mathbb{R}^{m\times n} = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} & \cdots & \frac{\partial f(A)}{\partial A_{1n}} \\ \frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots & \frac{\partial f(A)}{\partial A_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{m1}} & \frac{\partial f(A)}{\partial A_{m2}} & \cdots & \frac{\partial f(A)}{\partial A_{mn}} \end{bmatrix}$$

i.e., an $m \times n$ matrix with

$$(\nabla_A f(A))_{ij} = \frac{\partial f(A)}{\partial A_{ij}}.$$

The Gradient

Note that the size of $\nabla_A f(A)$ is always the same as the size of A. So if, in particular, A is just a vector $x \in \mathbb{R}^n$,

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}.$$

It follows directly from the equivalent properties of partial derivatives that:

- $\nabla_x (f(x) + g(x)) = \nabla_x f(x) + \nabla_x g(x)$.
- For $t \in \mathbb{R}$, $\nabla_x (t\, f(x)) = t \nabla_x f(x)$.

Matrix Calculus

The Hessian

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a function that takes a vector in $\mathbb{R}^n$ and returns a real number. Then the ***Hessian*** matrix with respect to x, written $\nabla_x^2 f(x)$ or simply as H is the $n \times n$ matrix of partial derivatives,

$$\nabla_x^2 f(x) \in \mathbb{R}^{n\times n} = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}.$$

In other words, $\nabla_x^2 f(x) \in \mathbb{R}^{n\times n}$, with

$$(\nabla_x^2 f(x))_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}.$$

Note that the Hessian is always symmetric, since

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x)}{\partial x_j \partial x_i}.$$

Gradients of Linear Functions

For $x \in \mathbb{R}^n$, let $f(x) = b^T x$ for some known vector $b \in \mathbb{R}^n$. Then

$$f(x) = \sum_{i=1}^{n} b_i x_i$$

so

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} b_i x_i = b_k.$$

From this we can easily see that $\nabla_x b^T x = b$. This should be compared to the analogous situation in single variable calculus, where $\partial/(\partial x)\ ax = a$.

Matrix Calculus

Gradients of Quadratic Function

Now consider the quadratic function $f(x) = x^T Ax$ for $\boxed{A \in \mathbb{S}^n.}$ Remember that

$$f(x) = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}x_i x_j.$$

To take the partial derivative, we'll consider the terms including x_k and x_k^2 factors separately:

$$\begin{aligned}
\frac{\partial f(x)}{\partial x_k} &= \frac{\partial}{\partial x_k}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}x_i x_j \\
&= \frac{\partial}{\partial x_k}\left[\sum_{i\neq k}\sum_{j\neq k} A_{ij}x_i x_j + \sum_{i\neq k} A_{ik}x_i x_k + \sum_{j\neq k} A_{kj}x_k x_j + A_{kk}x_k^2\right] \\
&= \sum_{i\neq k} A_{ik}x_i + \sum_{j\neq k} A_{kj}x_j + 2A_{kk}x_k \\
&= \sum_{i=1}^{n} A_{ik}x_i + \sum_{j=1}^{n} A_{kj}x_j = 2\sum_{i=1}^{n} A_{ki}x_i, \longrightarrow \boxed{\frac{\partial f(x)}{\partial x} = 2Ax}
\end{aligned}$$

Hessian of Quadratic Functions

Finally, let's look at the Hessian of the quadratic function $f(x) = x^T A x$
In this case,

$$\frac{\partial^2 f(x)}{\partial x_k \partial x_\ell} = \frac{\partial}{\partial x_k}\left[\frac{\partial f(x)}{\partial x_\ell}\right] = \frac{\partial}{\partial x_k}\left[2\sum_{i=1}^{n} A_{\ell i} x_i\right] = 2A_{\ell k} = 2A_{k\ell}.$$

Therefore, it should be clear that $\nabla_x^2 x^T A x = 2A$, which should be entirely expected (and again analogous to the single-variable fact that $\partial^2/(\partial x^2)\ ax^2 = 2a$).

Matrix Calculus Example: Least Squares

- Given a $\boxed{\text{full rank}}$ matrices $A \in \mathbb{R}^{m \times n}$, and a vector $b \in \mathbb{R}^m$ such that $b \notin \mathcal{R}(A)$, we want to find a vector x such that Ax is as close as possible to b, as measured by the square of the Euclidean norm $\|Ax - b\|_2^2$.

- Using the fact that $\|x\|_2^2 = x^T x$, we have

$$\|Ax - b\|_2^2 = (Ax - b)^T (Ax - b) = x^T A^T Ax - 2b^T Ax + b^T b$$

- Taking the gradient with respect to x we have:

$$\begin{aligned} \nabla_x (x^T A^T Ax - 2b^T Ax + b^T b) &= \nabla_x x^T A^T Ax - \nabla_x 2b^T Ax + \nabla_x b^T b \\ &= 2A^T Ax - 2A^T b \end{aligned}$$

- Setting this last expression equal to zero and solving for x gives the normal equations

$$x = \boxed{(A^T A)^{-1}} A^T b$$

Roadmap

- Course logistics
- Math review
  - Probability
  - Linear algebra
  - Optimization
  - Matrix calculus
- Python and Google Colab set up

Thank you!

Q & A