CS181 Section #0 Note, Condensed

Due: never

1 Linear Algebra

1.1 Scalars and Vectors

Scalar: A **scalar** is a single element of a field, e.g. 5.

Vector: A **vector** is an ordered collection of n coordinates, where each coordinate is a scalar of the underlying field.

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

Norms: The formula for the $\mathbf{L}\boldsymbol{n}$ norm is given by:

$$||\mathbf{x}||_n = \sqrt{\sum_{i=1}^{n} x_i^n}$$

Inner Product: Also called the **dot product** or **scalar product**, this is equal to:

$$\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i=1}^{n} u_i v_i = ||\mathbf{u}||_2 ||\mathbf{v}||_2 \cos \alpha$$

where α is the angle between $\mathbf{u}$ and $\mathbf{v}$. Note that: $\langle \mathbf{u}, \mathbf{u} \rangle = ||\mathbf{u}||_2^2$, since $\alpha = 0$.

1.2 Linear Independence

A set of non-zero vectors $\{\mathbf{v}_1, ..., \mathbf{v}_n\}$ is **linearly independent** if the equation $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 +, ..., +c_n\mathbf{v}_n = \mathbf{0}$ for scalars $c_1, ..., c_n$ can only be satisfied by setting $c_1, ..., c_n$ all to 0.

1.3 Spaces and Subspaces

Vector space: A **vector space** $\mathcal{V}$ is a collection of vectors that satisfy the following properties:

- Closure under scaling: $\forall \mathbf{v} \in \mathcal{V}$ and scalars a, $a\mathbf{v} \in \mathcal{V}$
- Closure under addition: $\forall \mathbf{u}, \mathbf{v} \in \mathcal{V}, (\mathbf{u} + \mathbf{v}) \in \mathcal{V}$

Orthonormal basis: The set of vectors $\{\mathbf{v}_1, ..., \mathbf{v}_n\}$ form an **orthonormal basis** for $\mathcal{V}$ if they are all unit vectors ("normal") and if $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0, \forall i \neq j$ ("orthogonal") where $\langle , \rangle$ is the inner product.

1.4 Scalar, Vector, and Subspace Projection

For vectors $\mathbf{u}, \mathbf{v} \in \mathcal{V}$ and $\mathbf{v} \neq \mathbf{0}$, the **scalar projection** a of $\mathbf{u}$ onto $\mathbf{v}$ is computed as:

$$a = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{||\mathbf{v}||}$$

Using this, the **vector projection p** of $\mathbf{u}$ onto $\mathbf{v}$ can be computed as:

$$a\left(\frac{1}{||\mathbf{v}||}\mathbf{v}\right) = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle}\mathbf{v}$$

The **subspace projection p** of $\mathbf{u}$ onto $\mathcal{S}$ can be expressed as the sum of the projections of $\mathbf{u}$ onto each element of the basis of $\mathcal{S}$:

$$\mathbf{p} = \sum_{i=1}^{m} \frac{\langle \mathbf{u}, \mathbf{s}_i \rangle}{\langle \mathbf{s}_i, \mathbf{s}_i \rangle}\mathbf{s}_i$$

1.5 Matrices

A **matrix** is a rectangular array of scalars. We write matrices in **bold uppercase**.

If we have $\mathbf{A} \in \mathbb{R}^{n \times m}$, then the matrix $\mathbf{A}$ is an $n \times m$ matrix that represents a **linear transformation** from m to n dimensions, where $\mathbf{A}$ is an **operator**. A_{ij} is the scalar found at the i^{th} row and j^{th} column.

$$\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1m} \\ \vdots & \ddots & & \vdots \\ A_{n1} & A_{n2} & \dots & A_{nm} \end{bmatrix}$$

A typical linear transformation looks like the following, where $\mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{n \times m}$:

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

1.6 Matrix Properties

- $\mathbf{A}^\top$ is the **transpose** of $\mathbf{A}$ and has $A^\top_{ji} = A_{ij}$.
- $\mathbf{A}$ is **symmetric** if $A_{ij} = A_{ji}$. That is, $\mathbf{A} = \mathbf{A}^\top$. Only square matrices can be symmetric.
- $\mathbf{A}$ is **orthogonal** if its rows and columns are orthogonal unit vectors. Consequence: $\mathbf{A}^\top\mathbf{A} = \mathbf{A}\mathbf{A}^\top = \mathbf{I}$ where $\mathbf{I}$ is the **identity matrix** (ones on the main diagonal and zeros elsewhere). Orthogonal matrix $\mathbf{A}$ has $\mathbf{A}^\top = \mathbf{A}^{-1}$.
- **Diagonal** matrices have non-zero values on the main diagonal and zeros elsewhere.
- **Upper-triangular** matrices only have non-zero values on the diagonal or above (top right of matrix).
- **Lower-triangular** matrices only have non-zero values on the diagonal or below (bottom right of matrix).

1.7 Matrix Multiplication

$\mathbf{AB}$ is a valid **matrix product** if $\mathbf{A}$ is $p \times q$ and $\mathbf{B}$ is $q \times r$ (left matrix has same number of columns as right matrix has rows).

Properties of matrix multiplication:

- $\mathbf{AB} \neq \mathbf{BA}$ (usually)
- $\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB}+\mathbf{AC}$ and $(\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC}$.
- $\lambda(\mathbf{AB}) = (\lambda\mathbf{A})\mathbf{B}$ and $(\mathbf{AB})\lambda = \mathbf{A}(\mathbf{B}\lambda)$, for some scalar λ.
- $(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top$

1.8 Rank, Determinant, Inverse

Rank: The **rank** of a matrix is the **dimension** of the vector space spanned by its column vectors. A matrix is full rank if all its column vectors are linearly independent.

Determinant: The **determinant** of a square matrix is a scalar quantity. $det(\mathbf{A})$ is equal to the product of the eigenvalues of $\mathbf{A}$. *Note:* You may also see the determinant denoted with single bars, e.g. $|\mathbf{X}|$.

Inverse: The **inverse** $\mathbf{A}^{-1}$ "undoes" $\mathbf{A}$ much like multiplying by $\frac{1}{x}$ undoes multiplying by x. $\mathbf{A}^{-1}$ only exists if $det(\mathbf{A}) \neq 0$. It is a given that $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$.

Moore-Penrose Pseudoinverse: The **Moore-Penrose pseudoinverse** $\mathbf{A}^+$ of $\mathbf{A}$ is a generalization of the inverse to non-square matrices, where $\mathbf{AA}^+\mathbf{A} = \mathbf{A}$. However, $\mathbf{AA}^+$ may not be the general identity matrix but maps all column vectors of $\mathbf{A}$ to themselves.

1.9 Eigen-Everything

Eigenvalues: If $\mathbf{Ax} = \lambda\mathbf{x}$ for some scalar λ, then λ is an **eigenvalue** of $\mathbf{A}$ and $\mathbf{x}$ is an **eigenvector**.

Eigen-decomposition: Let $\mathbf{A}$ be an $n \times n$ full-rank matrix with n linearly independent eigenvectors $\{\mathbf{q}_i\}_{i=1}^n$. $\mathbf{A}$ can be factored into $\mathbf{A} = \mathbf{Q\Lambda Q}^{-1}$ where $\mathbf{Q}$ is $n \times n$ and has $\mathbf{q}_i$ for its i^{th} column. $\mathbf{\Lambda}$ is a diagonal matrix whose elements are the corresponding eigenvalues: $\Lambda_{ii} = \lambda_i$. If a $\mathbf{A}$ can be eigen-decomposed and none of its eigenvalues are 0, then $\mathbf{A}$ is **nonsingular** and its inverse is given by $\mathbf{A}^{-1} = \mathbf{Q\Lambda}^{-1}\mathbf{Q}^{-1}$ with $\mathbf{\Lambda}_{ii}^{-1} = \frac{1}{\lambda_i}$.

Singular Value Decomposition: Generalizes eigen-decomposition to rectangular matrices. Let $\mathbf{A}$ be an $m \times n$ matrix. Then $\mathbf{A}$ can be factored into $\mathbf{U\Sigma V}^\top = \mathbf{U\Sigma V}^{-1}$ where

- $\mathbf{U}$ is $m \times m$ and orthogonal. The columns of $\mathbf{U}$ are the **left-singular vectors** of $\mathbf{A}$.
- $\mathbf{\Sigma}$ is an $m \times n$ diagonal matrix with non-negative real entries. The diagonal values σ_i of $\mathbf{\Sigma}$ are known as the **singular values** of $\mathbf{A}$. These are also the square roots of the eigenvalues of $\mathbf{A}^\top\mathbf{A}$.
- $\mathbf{V}$ is an $n \times n$ orthogonal matrix. The columns of $\mathbf{V}$ are the **right-singular** vectors of $\mathbf{A}$.

1.10 Positive Definiteness

Positive definite: Symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is **positive definite** if, for all non-zero vector $\mathbf{x} \in \mathbb{R}^n$:

$$\mathbf{x}^\top\mathbf{Ax} > 0$$

Positive semi-definite: Symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is **positive semi-definite** if, for all non-zero vector $\mathbf{x} \in \mathbb{R}^n$:

$$\mathbf{x}^\top \mathbf{A} \mathbf{x} \geq 0$$

Positive definite means all eigenvalues > 0, while **positive semi-definite** means all eigenvalues ≥ 0.

2 Calculus

2.1 Differentiation

$$\begin{aligned}
\text{Chain rule: } & \frac{d}{dx} f(g(x)) = f'(g(x))g'(x) \\
\text{Product rule: } & \frac{d}{dx} f(x)g(x) = f'(x)g(x) + f(x)g'(x) \\
\text{Linearity: } & \frac{d}{dx}(af(x) + bg(x)) = af'(x) + bg'(x)
\end{aligned}$$

for scalars a and b.

The **Jacobian** is a matrix where the j^{th} column is made up of the partial derivatives of f_j (the j^{th} output value of $\mathbf{f}$) with respect to all input elements, rows $i = 1$ to n.

$$\frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1(\mathbf{x})}{\partial x_n} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

If f is scalar-valued, its derivative is a column vector we call the **gradient vector**:

$$\frac{df(\mathbf{x})}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \cdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

The gradient vector points in the direction of steepest ascent in $f(\mathbf{x})$, which is useful for optimization.

A few important derivatives:

$$\begin{aligned}
\frac{\partial \mathbf{x}^\top \mathbf{a}}{\partial \mathbf{x}} &= \frac{\partial \mathbf{a}^\top \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a} \\
\frac{\partial \mathbf{a}^\top \mathbf{X} \mathbf{b}}{\partial \mathbf{X}} &= \mathbf{a}\mathbf{b}^\top \\
\frac{\partial (\mathbf{x} - \mathbf{A}\mathbf{s})^T \mathbf{W} (\mathbf{x} - \mathbf{A}\mathbf{s})}{\partial \mathbf{s}} &= -2\mathbf{A}^T \mathbf{W} (\mathbf{x} - \mathbf{A}\mathbf{s}) \\
\frac{\partial \mathbf{a}^\top \mathbf{X}^\top \mathbf{b}}{\partial \mathbf{X}} &= \mathbf{b}\mathbf{a}^\top \\
\frac{\partial \mathbf{a}^\top \mathbf{X} \mathbf{a}}{\partial \mathbf{X}} &= \frac{\partial \mathbf{a}^\top \mathbf{X}^\top \mathbf{a}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{a}^\top \\
\frac{\partial \mathbf{X}}{\partial X_{ij}} &= \mathbf{J}^{ij} \quad ***
\end{aligned}$$

*** $\mathbf{J}$ is NOT the Jacobian, but rather, a matrix with all zeros except for a 1 in the i, j entry.

For more matrix derivatives, see the **Matrix Cookbook** linked on the course website.

2.2 Optimization

Local Extrema: The local extrema of a single-variable function can be found by solving $\frac{d\mathbf{f}(\mathbf{x})}{d\mathbf{x}} = \mathbf{0}$. However, this equation is often intractable. We can search for local minima numerically using gradient-based methods.

Gradient Descent: Start with an initial guess $\mathbf{w}_0$ for the value of parameter $\mathbf{w}$. At each step i, update our guess for $\mathbf{w}$ by going in the direction of greatest descent of a loss function (opposite the gradient vector):

$$\mathbf{w}_{i+1} = \mathbf{w}_i - \eta \frac{df(\mathbf{w})}{d\mathbf{w}}$$

where η is a learning rate. We stop when the value of the gradient is close to 0.

3 Probability Theory

3.1 Random Variables

Discrete: Takes a value from a sample space $\mathcal{X}$ of discrete values. $p(x)$ is the **probability mass function** of X and can also be written as $p_X(x)$. We say that $x \sim X$ (x is sampled from X) when the value of x is picked in accordance with the distribution of X.

Continuous: Can take on a continuous range of values. $p(x)$ or $p_X(x)$ represents the **probability density function** of a continuous random variable. The probability of any one exact value is zero.

3.2 Expectation

The **expected value** (or *expectation* or *mean*) of a random variable can be thought of as the "weighted average" of the possible outcomes of the random variable. For discrete variables:

$$\mathbb{E}_{x \sim p(x)}[X] = \sum_{x \in \mathcal{X}} x \cdot p(x) \qquad \mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} f(x)p(x)$$

For continuous variables:

$$\mathbb{E}[X] = \int_{\mathcal{X}} x \cdot p(x)dx \qquad \mathbb{E}[f(X)] = \int_{\mathcal{X}} f(x)p(x)dx$$

Properties of expectation:

- $\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$
- $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ if X and Y are independent

3.3 Variance

Variance is a measure of the spread of a random variable.

$$\begin{aligned} var(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \end{aligned}$$

Properties of variance:

$$var(aX + b) = a^2 var(X)$$

3.4 Joint Probability

The **joint probability** of $X = x$ and $Y = y$ is written as $p(x, y)$ or $p_{XY}(x, y)$.

If X and Y are **independent**, then: $p(x, y) = p(x)p(y)$.

It will **always be true** that: $p(x, y) = p(x)p(y|x) = p(y)p(x|y)$

Convert a joint probability $p(x, y)$ to the **marginal distribution** of a single variable, e.g. $p(x)$, by summing:

$$\text{Discrete: } p(x) = \sum_{y \in \mathcal{Y}} p(x, y) \qquad \text{Continuous: } p(x) = \int_{y \in \mathcal{Y}} p(x, y)$$

3.5 Conditional Probability

$X|Y$ represents the random variable X conditioned on the random variable Y.

$$p(x|y) = \frac{p(x, y)}{p(y)}$$

We can factor a joint probability into chains of conditional probabilities with the **product rule**:

$$\begin{aligned} p(x, y, z) &= p(x)p(y|x)p(z|x, y) \\ &= p(y)p(x|y)p(z|x, y) \\ &= p(z)p(x|z)p(y|x, z) \\ &= \text{etc...} \end{aligned}$$

3.6 Bayes' Theorem

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Since we are conditioning on y, that means y is constant and can be replaced with a normalizing constant:

$$p(x|y) \propto p(y|x)p(x)$$

3.7 Covariance

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Properties of covariance: (supposing X, Y, Z have mean 0 and finite variances)

- Symmetric: $\text{cov}(X, Y) = \text{cov}(Y, X)$
- Positive Semi-definite: $\text{cov}(X, X) \geq 0$
- Bilinear: $\text{cov}(aX + bY, Z) = a\text{cov}(X, Z) + b\text{cov}(Y, Z)$

The $n \times n$ **covariance matrix** (often denoted $\mathbf{\Sigma}$), where $\mathbf{\Sigma}_{ij} = \text{cov}(X_i, X_j)$ is the empirical covariance between the i^{th} and j^{th} features.

3.8 Conditional Expectation and Conditional Variance

The **conditional expectation** of X given $Y = y$ is: $\mathbb{E}[X|Y]$.

Similarly, **conditional variance** is: $var(X|Y) = \mathbb{E}[(X - \mathbb{E}[X|Y])^2|Y] = \mathbb{E}[X^2|Y] - \mathbb{E}[X|Y]^2$

Properties:

- $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$
- $\mathrm{var}[X] = \mathbb{E}[\mathrm{var}[X|Y]] + \mathrm{var}[\mathbb{E}[X|Y]]$

3.9 Gaussians

3.9.1 Univariate PDF

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

- If X, Y are independent normals then $X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$
- $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$
- Any PDF proportional to $\exp(ax^2 + bx + c)$ must be a Gaussian PDF.

3.9.2 Multivariate PDF

Given dimension m, mean vector $\mu \in \mathbb{R}^m$, and covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{m \times m}$,

$$\mathcal{N}(\mathbf{x}; \mu, \mathbf{\Sigma}) = \frac{1}{det(2\pi\mathbf{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu)\right)$$