CS 181 Spring 2016 Section 1 Notes (Math Review)

1 Probability

1.1 Random Variables

A random variable X is a variable which could take on various values on the real line with specified probabilities (its distribution). An event A is something that could happen (e.g. $X = x$, the event that X takes on the value of a number x). $p(A)$ is the probability that A happens. Following Bishop, we won't be too strict about only writing events as arguments to the probability function p. We will use $p(X)$ to mean the distribution of X, and $p(x)$ to mean $p(X = x)$.

1.2 Expected Value

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

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

For continuous random variables:

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

The most important property of expected values is the **linearity of expectation**. For **any** two random variables X and Y, a and b scaling coefficients and c is our constant, the following property of holds:

$$\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$$

The above is true regardless of whether X and Y are dependent or independent.

1.3 Variance

The variance of a random variable is its expected squared deviation from its mean

$$\text{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

1.4 Conditional Probability

When we know that an event B has happened, that could change the probability of another event A. The new probability of A given B is the conditional probability $p(A|B)$. If B changes the distribution of a random variable X, we write the new random variable as $X|B$, and the new distribution $p(X|B)$ is the conditional distribution.
Note that $\mathbb{E}[X|Y]$ is a random variable (this is one form of $f(Y)$). Adam's law (law of iterated expectations) gives

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$$

There is an analogous property for variances (Eve's Law, or law of total variance)

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

1.5 Sum and Product Rules

The sum rule allows us to find the marginal probability $p(x)$:

$$p(x) = \sum_{y \in \mathcal{Y}} p(x, y)$$
$$p(x) = \int_{\mathcal{Y}} p(x, y) dy$$

The product rule gives the joint probability $p(x, y)$ as the product of a conditional probability and a marginal probability:

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

which can be extended to as many variables as you want

$$\begin{aligned} p(x_1, ..., x_n) &= p(x_1|x_2, ..., x_n)p(x_2, ..., x_n) \\ &= ... = p(x_1|x_2, ..., x_n)...p(x_{n-1}|x_n)p(x_n) \end{aligned}$$

1.6 Bayes' Theorem

$$\begin{aligned} p(x|y) &= \frac{p(y|x)p(x)}{p(y)} \\ &= \frac{p(y|x)p(x)}{\int_{\mathcal{X}} p(x, y) dx} \end{aligned}$$

Bayesian interpretation: say we observe data y, and we are interested in parameters θ. We can write the *posterior distribution* of the parameters given data by by using Bayes' theorem.

$$\underbrace{p(\theta|y)}_{\text{posterior}} = \frac{\overbrace{p(y|\theta)}^{\text{likelihood}} \overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(y)}_{\text{evidence}}}$$

1.7 Distributions

A distribution describes the probability that a random variable takes on certain values. The range of values that the random variable could take on is the support of its distribution.
Here are the properties of some famous distributions:

Univariate Distributions

Distribution	**Support and PDF**	$\mathbb{E}[x]$	$\text{var}[x]$
Bernoulli	$x \in \{0, 1\}$ $\text{Bern}(x\|\mu) = \mu^x(1-\mu)^{1-x}$	μ	$\mu(1-\mu)$
Uniform	$x \in (a, b)$ $\text{U}(x\|a, b) = \frac{1}{b-a}$	$\frac{b+a}{2}$	$\frac{(b-a)^2}{12}$
Beta	$x \in (0, 1)$ $\text{Beta}(x\|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$
Binomial	$x \in \{0, 1, ..., N\}$ $\text{Bin}(x\|N, \mu) = \binom{N}{x}\mu^x(1-\mu)^{N-x}$	$N\mu$	$N\mu(1-\mu)$
Gamma	$x > 0$ $\text{Gam}(x\|a, b) = \frac{1}{\Gamma(a)}b^a x^{a-1}e^{-bx}$	$\frac{a}{b}$	$\frac{a}{b^2}$
Univariate Gaussian	$x \in \mathbb{R}$ $\mathcal{N}(x\|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$	μ	σ^2

Multivariate Distributions

Distribution	**Support and PDF**	$\mathbb{E}[x_k]$	$\text{var}[x_k]$	$\text{cov}[x_j, x_k], j \neq k$
Categorical	$x_k \in \{0, 1\}, \sum_{k=1}^K x_k = 1$ $\text{Cat}(\mathbf{x}\|\boldsymbol{\mu}) = \prod_{k=1}^K \mu_k^{x_k}$	μ_k	$\mu_k(1-\mu_k)$	$-\mu_j\mu_k$
Multinomial	$x_k \in \{0, 1, ..., N\}, \sum_{k=1}^K x_k = N$ $\text{Mult}(x_1, x_2, \ldots, x_K\|\boldsymbol{\mu}, N) = \binom{N}{x_1 x_2 \ldots x_K}\prod_{k=1}^K \mu_k^{x_k}$	$N\mu_k$	$N\mu_k(1-\mu_k)$	$-N\mu_j\mu_k$
Multivariate Gaussian	$\mathbf{x} \in \mathbb{R}^D$ $\mathcal{N}(\mathbf{x}\|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}\|\boldsymbol{\Sigma}\|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$	μ_k	Σ_{kk}	Σ_{jk}

Aside: In machine learning, statistical distributions are useful for modelling how our data is generated. We then make use of these statistical models for tasks such as prediction and classification.

2 Linear Algebra

2.1 Vectors

$$\mathbf{x} = (x_1, \ldots, x_D)^\top$$
$$\mathbf{x} \in \mathbb{R}^D$$

where D is the dimension of the vector, and $x_1, \ldots, x_D$ are elements of the vector.
We follow Bishop and use column vectors by default.

2.2 Matrices

Technically, a $N \times M$ matrix is a (linear) transformation from $\mathbb{R}^N$ to $\mathbb{R}^M$. This matrix has N rows and M columns. If we call the matrix $\mathbf{A}$, the element at the n^{th} row and m^{th} column of $\mathbf{A}$ is A_{nm}.

The transpose of a matrix $\mathbf{A}$ switches the rows and columns, i.e. if $\mathbf{B} = \mathbf{A}^\top$, then $B_{mn} = A_{nm}$.
A matrix $\mathbf{A}$ is symmetric if $A_{nm} = A_{mn}$. Only square matrices can be symmetric.

The inverse of a matrix $\mathbf{A}$ is the unique matrix $\mathbf{A}^{-1}$ such that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, where $\mathbf{I}$ is identity matrix. The inverse only exists for square matrices. If $\mathbf{A}$ is a rectangular matrix, we can define the Moore-Penrose pseudoinverse $\mathbf{B} = \left(\mathbf{A}^\top\mathbf{A}\right)^{-1}\mathbf{A}^\top$, such that $\mathbf{B}\mathbf{A} = \mathbf{I}$, where $\mathbf{I}$ is the $M \times M$ identity matrix if $\mathbf{A}$ is a $N \times M$ matrix.

A matrix $\mathbf{A}$ is orthogonal if $\mathbf{A}^\top = \mathbf{A}^{-1}$. These matrices preserve inner products, and represent an orthogonal change of basis (i.e. rotations and reflections).

Aside: In machine learning, matrices often come up in the context of covariances between input features. We also often construct *design matrices*, where each row in the matrix is an input from our dataset, to simplify computation by using matrix operations.

2.3 Solutions to Linear Systems

A system of linear equations with N equations and $M - 1$ variables can be written as a matrix equation

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$
$$\mathbf{y} \in \mathbb{R}^N, \mathbf{A} \in \mathbb{R}^{N \times M}, \mathbf{x} \in \mathbb{R}^M$$

where x has M dimensions because of constants. We can put $\mathbf{A}$ into reduced row echelon form (rref) through a series of row operations that involve only

- Switching two rows

- Multiplying a row by a constant
- Adding one row to another

You can convince yourself that these are all legitimate procedures to carry out when solving simultaneous equations. These operations can be carried out by left-multiplying $\mathbf{A}$ with what are called elementary matrices. Suppose the sequence $\mathbf{E_n} \dots \mathbf{E_1}$ puts $\mathbf{A}$ into rref (i.e. the matrix $\mathbf{E_n} \dots \mathbf{E_1 A}$ is in rref), then our original equation becomes

$$\mathbf{E_n} ... \mathbf{E_1 y} = \mathbf{E_n} ... \mathbf{E_1 A x}$$

After calculating $\mathbf{E_n} ... \mathbf{E_1 y}$, we can easily read off the solutions (or lack thereof) to the linear system of equations.

2.4 Subspaces

Linear subspace: given a set of vectors $\mathbf{x}_1, \dots, \mathbf{x}_D$ (we can assume these are linearly independent without loss of generality), where $\mathbf{x}_d \in \mathbb{R}^N$, and $D \leq N$, we say these vectors define a linear space if we look at the set of points

$$S = \{y : y = \mathbf{w}^\intercal \mathbf{X}, \text{ for all } \mathbf{w} \in \mathbb{R}^D\}$$

This is the space that is *spanned* by the vectors $\mathbf{x}_1, \dots, \mathbf{x}_D$. If $D < N$, then S is a proper subspace of $\mathbb{R}^N$.

2.5 Projection

If a subspace S is spanned by vectors $\mathbf{x}_1, \dots, \mathbf{x}_D$ (we can choose these to be orthonormal without loss of generality) where $\mathbf{x}_d \in \mathbb{R}^N$, then the projection matrix is the design matrix $\mathbf{X}$, so that for any $\mathbf{y} \in \mathbb{R}^N$, $\mathbf{Xy}$ is in subspace S, with the components of $\mathbf{y}$ projected onto S (components of $\mathbf{y}$ that are orthogonal to S are "discarded").

2.6 Matrix Rank

One definition: the dimensionality of the span of the matrix (column or row). It's the number of independent column or row vectors in the matrix.

2.7 Matrix Determinant

The general formula is quite complicated, but you should know that $\det(\mathbf{A}) = 0$ if and only if $\mathbf{A}$ is singular (i.e. has no inverse). This is due to the property that $\det(\mathbf{A}^{-1}) = 1/\det(\mathbf{A})$, which is undefined if $\det(\mathbf{A}) = 0$.

2.8 Positive Definite Matrices

A symmetric matrix $A \in \mathbb{R}^{N \times N}$ is **positive definite** if it satisfies the property

$$\mathbf{y}^\intercal A \mathbf{y} > 0$$

and positive semi-definite if it satisfies

$$\mathbf{y}^\intercal A \mathbf{y} \geq 0$$

for every non-zero vector $\mathbf{y} \in \mathbb{R}^N$.

2.9 Eigenvalues and Eigenvectors

A matrix $\mathbf{A}$ is said to have eigenvector $\mathbf{v}$ with eigenvalue λ if we can write

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

i.e. the matrix only modifies the magnitude of the vector $\mathbf{v}$. We note that the matrix $\mathbf{A} - \lambda \mathbf{I}$ must be singular since $(\mathbf{A} - \lambda \mathbf{I})\mathbf{v} = 0$. This means that we can solve for eigenvalues by using the equation

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0$$

$\det(\mathbf{A} - \lambda \mathbf{I})$ is called the characteristic polynomial of $\mathbf{A}$.

2.10 Singular Value Decomposition (SVD)

For any symmetric matrix $\mathbf{A}$, we can always write

$$\mathbf{A} = \mathbf{R}\mathbf{V}\mathbf{R}^{-1}$$

where $\mathbf{R}$ is an orthogonal matrix and $\mathbf{V}$ is a diagonal matrix. Because orthogonal matrices represent orthogonal changes of basis, this means that we are finding a basis under which $\mathbf{A}$ is diagonal. You can convince yourself that the diagonal entries of $\mathbf{V}$ are the eigenvalues of $\mathbf{A}$.This procedure is known as diagonalizing the matrix, and $\mathbf{R}\mathbf{V}\mathbf{R}^{-1}$ is the singular value decomposition of $\mathbf{A}$.
If we are further given that $\mathbf{A}$ is positive definite, then its eigenvalues (i.e. the diagonal elements of $\mathbf{V}$) are all positive.

Aside: SVD is often used in machine learning to better understand a dataset by showing the number of important dimensions, as well as to get rid of redundant data through dimensionality reduction.

3 Multivariate Calculus

Often we are interested in functions $f : \mathcal{X} \to \mathcal{Y}$ and their derivatives to fit a model from observations. The most salient is the objective function

$$E_D(\mathbf{w}) = \sum_{n=1}^{N} (t_n - \mathbf{w}^\intercal \mathbf{x}_n)^2$$

This section will focus on functions from many inputs (e.g. $\mathcal{X} = \mathbb{R}^D$ for some large D). We will consider $\mathcal{Y} = \mathbb{R}^M$, but the case of $M = 1$ will be the most common and useful.

3.1 Differentiation

$$\text{Chain rule: } \frac{d}{dx} f(g(x)) = f'(g(x))g'(x)$$
$$\text{Product rule: } \frac{d}{dx} f(x)g(x) = f'(x)g(x) + f(x)g'(x)$$
$$\text{Quotient rule: } \frac{d}{dx} \frac{f(x)}{g(x)} = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$$

3.2 Gradient Vector

Gradient vector is with respect to function inputs:

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_D} \right)^\top$$

Sometimes we only wish to differentiate with respect to some variables in the input. For example, $f(\mathbf{x}, \alpha)$, then the derivative with respect to $\mathbf{x}$ is denoted

$$\nabla_{\mathbf{x}} f(\mathbf{x}, \alpha) = \frac{df(\mathbf{x}, \alpha)}{d\mathbf{x}} = \left(\frac{\partial f(\mathbf{x}, \alpha)}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x}, \alpha)}{\partial x_D} \right)^\top$$

The gradient vector points towards the direction of greatest ascent in $f(\mathbf{x})$ with respect to the parameters being differentiated.

3.3 Gradient Descent

The local minima of a function can be found using the condition $\nabla f(\mathbf{x}) = 0$. However, we may not always be able to solve this equation, or if $f(\mathbf{x})$ is complicated we may get a variety of local maxima and saddle points that are not our desired minimum. Gradient descent is a numerical way to minimize $f(\mathbf{x})$. We start with an initial guess $\mathbf{x}_0$, and then

at each step i we update our guess by going in the direction of greatest descent (opposite the direction of the gradient vector)

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \gamma \nabla f(\mathbf{x}_i)$$

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

Aside: Gradient descent is often used in machine learning when we're dealing with an objective function that is not easily optimizable. There are also variants such as stochastic gradient descent, which are simple extensions of gradient descent, but are very useful in certain machine learning applications.

3.4 Jacobian Matrix

This is a generalization of the gradient vector to functions that have multiple outputs, e.g. $\mathbf{f}(\mathbf{x})$. The Jacobian matrix is defined as

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

3.5 Hessian Matrix

For a function that has a single output $f(\mathbf{x})$, the Hessian matrix is the generalization of the second derivative

$$H(f(\mathbf{x})) = \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{bmatrix}$$

This matrix is symmetric and we can diagonalize it to find its eigenvalues. If all eigenvalues are positive, then we are at a local minimum. If all eigenvalues are negative, then we are at a local maximum. If there is a mix of positive and negative eigenvalues, then we are at a saddle point. If there are zero-eigenvalues, then there are directions in which $f(\mathbf{w})$ has no second order change. Besides allowing us to deduce the nature of turning points, the inverse of the Hessian matrix can also be used to aid gradient descent by using the update rule

$$\mathbf{x}_{i+1} = \mathbf{x}_i - [H(f(\mathbf{x}))]^{-1} \nabla f(\mathbf{x}_i)$$

3.6 Convexity/Concavity

A single output function $f(\mathbf{x})$ is convex if it satisfies Jensen's inequality

$$f\left(\sum_{n=1}^{N} a_n \mathbf{x}_n\right) \leq \sum_{n=1}^{N} a_n f\left(\mathbf{x}_n\right)$$

for all $\mathbf{x}_n \in \mathcal{X}$ and $\sum_{n=1}^{N} a_n = 1$. Similarly, $f(\mathbf{x})$ is concave if the sign of the inequality is flipped. $f(\mathbf{x})$ is called strongly convex/concave if the inequalities are converted to strict inequalities.
Strongly convex and concave functions have the benefit of having global minimums and maximums respectively, and so are easy to do gradient descent on.