CS 182 Fall 2018 Math Review Notes

1 Probability Review

A **random variable** X is a variable which could take on various values with specified probabilities (its distribution). The possible outcomes of X are described through **events**. For instance, an event A could be that X takes on the value 2, or that $X < 4$.

The probability that A happens is denoted as $P(A)$. For continuous random variables X, the probability that X attains some value x is described by a function p, called the **probability density function (PDF)**. Similarly, for discrete random variables X, it is called the **probability mass function (PMF)**. 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 $p(X = x)$.

Events can be constructed from other events. For instance, given two events A and B, we may be interested in the event $C = (A \ and \ B)$ or $C = A \cap B$ (the event that both events happen) or the event $D = (A \ or \ B)$ or $D = A \cup B$ (the event that either of the events happens).

Two events A and B are **independent** if the occurence of one does not influence the probability of the other. This can be expressed as $P(A \cap B) = P(A)P(B)$. Similarly, two random variables X and Y are independent if $p(x, y) = p(x)p(y)$.

The **expected value** (or **expectation, mean**) $\mathbb{E}[X]$ of a random variable X can be thought of as the "weighted average" of the possible outcomes of the random variable. It is often represented by μ.

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:

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

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

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

This is easy to show from the definition above, and is true regardless of whether X and Y are dependent or independent.

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

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

The **standard deviation** of a random variable is the square root of the variance:

$$\sigma(X) = \sqrt{\text{var}(X)}$$

Now consider probability distributions over multiple variables. Let X and Y be two random variables (for instance, each can correspond to the rolls of two regular 6-sided dice). Then, the probability of an event that depends on both X and Y (for example, the event that $X == 1$ and $Y == 2$) will be given by a value in the **joint probability distribution** $p(x, y)$.

The **marginal probability distribution** is the probability distribution of a subset of variables from a **joint probability distribution**. For example, the marginal probability distribution $p(y)$ can be found by summing across all values of x in $p(x, y)$:

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

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

When we know that an event B has happened, that could influence 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.

The conditional probability can be defined as:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Note that this has some direct implications: for instance, if A is the event $X == x$ and B is the event $Y == y$, we have an expression relating distributions (sometimes called the **product rule**):

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

which implies that

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$$

Remember recursion? If there are more variables in the distribution, the above statement can be recursively extended:

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

Finally, you may have heard of **Bayes' Theorem** (also known as **Bayes' Rule** or **Bayes' Law**), a very prominent statement that relates conditional probabilities between events. For any two events A and B such that $P(B) > 0$,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

As before, if we set events A to be $X == x$ and B to be $Y == y$, we get another common form of Bayes' Law:

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

In machine learning, we are often looking for parameters θ based on the distribution of data y. Then, we interpret $p(\theta)$ as the **prior** (the distribution of θ without knowing about y), $p(y)$ as the **evidence** (the overall probability of this data without considering our parameters), $p(y|\theta)$ as the **likelihood** (how likely are we to collect this data given the parameters?) and $p(\theta|y)$ as the **posterior** (the distribution of θ after knowing about y).

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

2 Linear Algebra

We write a **vector** as

$$\mathbf{x} = (x_1, \ldots, x_D)^\top$$

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

We may commonly write $\mathbf{x} \in \mathbb{R}^D$. The symbol "$\mathbb{R}$" refers to the space of real numbers, "$\mathbb{R}^D$" refers to the D-dimensional space of real numbers, and the symbol "$\in$" means "in", so this is a concise way to say that x is in the D-dimensional space of real numbers.

The size of a vector is measured with the vector norm. Usually, we are referring to the function:

$$||x||_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_D^2}$$

We call this the 2-norm or Ł^2-norm, and it is actually just one of many Ł^p norms. You may also see the 1-norm or the infinity-norm:

$$||x||_1 = |x_1| + |x_2| + \ldots |x_D|$$
$$||x||_\infty = \max(|x_1|, |x_2|, \ldots, |x_D|)$$

An $n \times m$ **matrix A** has n rows and m columns:

$$\begin{pmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,m} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \ldots & a_{n,m} \end{pmatrix}$$

An $n \times m$ matrix can be thought of as a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$, and, similarly to the vectors, we write $A \in \mathbb{R}^{n \times m}$.

3 Multivariate Calculus

To review mathematical notation, we may write a function f as

$$f : \mathcal{X} \in R^n \to \mathcal{Y} \in R^m$$

This means that the function's **domain**, or set of legals inputs, is a set $\mathcal{X}$ of real-valued n-dimensional vectors, and the function returns a real-valued m-dimensional vector in the output set (or **range**) $\mathcal{Y}$.

Recall the rules for **differentiation** of a function with respect to one variable:

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

When calculating partial derivatives of a function of multiple variables, we write the **gradient vector** as

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

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.

If the function we are interested in has multiple outputs (it maps a vector $[x_1, \ldots, x_m]$ to a vector $[f_1, \ldots, f_n]$), we can generalize the gradient vector to the **Jacobian matrix**:

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

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

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

To find the local minima (or maxima) of a function $f(\mathbf{x})$, we can set the gradient equal to zero: $\nabla f(\mathbf{x}) = 0$. However, we may not always be able to find a closed-form solution. **Gradient descent** is a numerical way to solve this problem. We start with an initial guess $\mathbf{x}_0$, and iteratively take small steps in the direction of greatest descent until we reach a point where the gradient is close to zero. Since this direction is exactly opposite the direction of the gradient vector, the "update" of our guess at each iteration is

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

where γ determines our step size. This algorithm forms the basis for much of optimization – many of the common algorithms used within machine learning and other fields are variants and extensions of gradient descent.