PRELIMINARIES

STA414/2104 Winter 2021

University of Toronto

1. Some definitions.

- **Functions.** We use $f : \mathbb{R}^d \to \mathbb{R}$ to denote that f is a function, its argument is in $\mathbb{R}^d$ and its output is real valued (or in $\mathbb{R}$). We denote by $\nabla f(x) \in \mathbb{R}^d$, its gradient (See Section 3 for definition).
- **ℓ_p-norms.** Since we are mostly dealing with vectors in machine learning, we will use different norms a lot. Euclidean norm, denoted as $\|\cdot\|_2$, is the most commonly used norm. But we can also use ℓ_p norms which are defined as

$$\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p\right)^{1/p}. \tag{1.1}$$

  When we drop the subscript and use $\|\cdot\|$, this typically means the Euclidean norm $\|\cdot\|_2$.
- **Indicator function.** The indicator function is defined as $\delta(a, b) = 1$ if $a = b$, and $\delta(a, b) = 0$ if $a \neq b$. For example, $\delta(2, 2) = 1$ and $\delta(2, 2.1) = 0$.
- **argmin & argmax.** Assume that we are trying to find the point in $\mathbb{R}^d$ that minimizes $f(x)$. This point is denoted by $x_* = \text{argmin}_x f(x)$. In general, there can be many points that minimize the function $f(x)$. If this is the case, argmin function returns a set of minimizers, and notation is slightly different $x_* \in \text{argmin}_x f(x)$. The function argmax is similar. For example, given a vector $a \in \mathbb{R}^d$, let $f(x) = \|x - a\|_2$ be a function. Then,

$$a = \operatorname*{argmin}_x f(x). \tag{1.2}$$

  Another example is that we have a binary classification problem over $\{0, 1\}$ and we are using a decision tree to solve it. We want to predict the class assignment of a region with 5 samples $t_1 = 1, t_2 = 1, t_3 = 0, t_4 = 0, t_5 = 1$. Then the majority assignment based on these samples can be found by

$$\operatorname*{argmax}_{t \in \{0,1\}} \sum_{i=1}^{5} \delta(t, t_i) = 1, \tag{1.3}$$

  because the summation is equal to 3 when $t = 1$, and 2 when $t = 0$.

2. Random variables and vectors. Random vectors are simply vectors with each coordinate a random variable. You can think of a d-dimensional random vector $X \in \mathbb{R}^d$ as a d random variables X_i's stacked together to make a vector. Most probability rules we have for random variables also hold for random vectors.

- **Density**. If we have a random vector X and its each coordinate is a continuous random variable, then we can talk about probability density function $p(x) : \mathbb{R}^d \to \mathbb{R}$ associated with the random vector X. This is defined similar to the one variable case and for a set $A \subset \mathbb{R}^d$, it satisfies $\mathbb{P}(X \in A) = \int_A p(x)dx$.
  For example, if we have the multivariate Gaussian random variable with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d\times d}$, its density is given as

$$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}. \tag{2.1}$$

  See below for definitions of mean and variance in the multivariable setting.
  If we have two random vectors $X, Y \in \mathbb{R}^d$ with joint density $p(x, y)$, the standard rules of conditional density apply.

$$X|Y \sim p(x|y) = \frac{p(x, y)}{p(y)}. \tag{2.2}$$

- **Expectation.** If $X \in \mathbb{R}^d$ is a random vector, its expectation

$$\mathbb{E}[X] = \mu \in \mathbb{R}^d$$

  is also a vector and it is defined as $\mathbb{E}[X_i] = \mu_i$. Below are some properties.
  - For random vectors $X, Y \in \mathbb{R}^d$, and a constant matrix $A \in \mathbb{R}^{d\times d}$, we have

$$\mathbb{E}[X + AY] = \mathbb{E}[X] + A\mathbb{E}[Y].$$

  - If we have $X_1, X_2, ..., X_n$ random vectors with $\mathbb{E}X_i = \mu$, their sample mean has the expectation μ, i.e.,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mu.$$

    Note that independence is not required for the above result.

- **Conditional expectation.** For two random vectors X, Y with joint distribution $p(x, y)$, the conditional expectation of $X|Y$ is given by

$$\mathbb{E}[X|Y = y] = \int xp(x|y)dx.$$

  This is like fixing the value of the random variable Y, and taking expectation of X after. Note that in the conditional expectation, we integrate out the variable x, but the variable we condition on is not integrated. This is why, $E[X|Y = y]$ is a function of y.
  A rule that comes up in bias-variance decomposition is the law of iterated expectation:

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

  This can be easily shown using the properties of the density. That is,

$$\mathbb{E}[\mathbb{E}[X|Y]] = \int\left[\int xp(x|y)dx\right]p(y)dy = \int\int xp(x, y)dxdy = \mathbb{E}[X].$$

- **Variance.** Variance of a random vector $X \in \mathbb{R}^{d\times d}$ is defined as
$$\mathrm{Var}(X) = \mathbb{E}[(X-\mu)(X-\mu)^T] \in \mathbb{R}^{d\times d}.$$
Observe that variance of a random vector is a $d \times d$ matrix and its ij-th entry is given by $\mathrm{Var}(X)_{ij} = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] \in \mathbb{R}$.
- **Covariance.** Let X, Y be two random vectors in $\mathbb{R}^d$. Then their covariance is given as
$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_y)^T] \in \mathbb{R}^{d\times d}$$
where μ_X and μ_Y denotes the mean of X and Y respectively.
Now assume that X and Y are independent.
  (a) Their covariance is zero: $\mathbb{E}[(X-\mu_X)(Y-\mu_y)^T] = \mathbb{E}[(X-\mu_X)]\mathbb{E}[(Y-\mu_y)^T] = 0$.
  (b) For a constant matrix $A \in \mathbb{R}^{d\times d}$, we have
$$\mathrm{Var}(X + AY) = \mathrm{Var}(X) + A\mathrm{Var}(Y)A^T.$$
Next, assume that we have $X_1, X_2, ..., X_n$ independent random vectors with mean μ and covariance matrix Σ. Contrary to sample mean above, in this case independence is required.
  (c) Using the above, if $X \sim \mathcal{N}(\mu, \Sigma)$, then $AX \sim \mathcal{N}(A\mu, A\Sigma A^T)$. This follows from the fact that linear transformation of a Gaussian random vector is again Gaussian – a property also used in bivariate Gaussian distributions.

2.1. *Maximum likelihood estimator.* Assume that we observe N i.i.d. random vectors $\mathcal{D} = \{X_1, X_2, ..., X_N\}$ from a distribution $p(x|\theta)$. We assume that the distribution function $p(x|\theta)$ is known, but the parameter θ is not known. For example, distribution can be Gaussian, but its mean and variance is unknown.

Using the independence assumption, we write the joint density of N i.i.d. random vectors.
$$p(x_1, x_2, ..., x_n|\theta) = \prod_{i=1}^{N} p(x_i|\theta).$$
The main idea behind the maximum likelihood estimation (MLE) is that we plug in our observations $\mathcal{D}$ into the joint density and find the θ value that maximizes this likelihood function. That is,
$$\theta^{\mathrm{MLE}} = \operatorname*{argmax}_\theta \prod_{i=1}^{N} p(X_i|\theta).$$
When finding the maximum of a function, we take its derivative and set it equal to 0. However, derivative of products is not easy to handle; therefore, we first apply the log function which doesn't change the point where the maximum is attained. This transforms the product to a summation. That is,
$$\begin{aligned}\theta^{\mathrm{MLE}} &= \operatorname*{argmax}_\theta \prod_{i=1}^{N} p(X_i|\theta), = \operatorname*{argmax}_\theta \log\left(\prod_{i=1}^{N} p(X_i|\theta)\right),\\ &= \operatorname*{argmax}_\theta \sum_{i=1}^{N} \log p(X_i|\theta).\end{aligned}$$

Now, we can easily take derivatives and find the maximizer.

2.2. *Maximum A posteriori Probability.* Assume the previous setup that we observe N i.i.d. random vectors $\mathcal{D} = \{X_1, X_2, ..., X_N\}$ from a distribution $p(x|\theta)$. This time, we will also assume that θ is a random vector and its prior distribution is given by $p(\theta)$. This means that instead of treating the parameter θ as a constant as in MLE, we assume some prior knowledge on θ which comes from $p(\theta)$. Maximum A posteriori Probability (MAP) estimator of θ maximizes the posterior distribution $p(\theta|\text{data})$ obtained by the Bayes rule

$$p(\theta|\text{data}) = \frac{p(\text{data}|\theta)p(\theta)}{p(\text{data})}.$$

Therefore, MAP estimator is given by

$$\theta^{\text{MAP}} = \operatorname*{argmax}_\theta p(\theta|\text{data}) = \operatorname*{argmax}_\theta p(\text{data}|\theta)p(\theta).$$

In the above maximization problem, we dropped the term in the denominator since it doesn't have the optimization parameter θ in it and doesn't contribute to the minimization problem.

For example in the previous setup, MAP estimator can be written as

$$\begin{aligned}\theta^{\text{MAP}} &= \operatorname*{argmax}_\theta p(X_1, .., X_N|\theta)p(\theta) = \operatorname*{argmax}_\theta p(\theta)\prod_{i=1}^N p(X_i|\theta),\\ &= \operatorname*{argmax}_\theta \log p(\theta) + \sum_{i=1}^N \log p(X_i|\theta).\end{aligned}$$

3. Basic multivariable calculus. For a given function $f : \mathbb{R}^d \to \mathbb{R}$, we denote its partial derivative with respect to its i-th coordinate as $\partial f(x)/\partial x_i \in \mathbb{R}$. Gradient of this function is simply a vector with i-th coordinate $\partial f(x)/\partial x_i \in \mathbb{R}$. That is,

$$[\nabla f(x)]_i = \frac{\partial f(x)}{\partial x_i}. \tag{3.1}$$

The gradient of a function points in the direction of greatest increase, and its magnitude is the rate of increase in that direction. Therefore, when you are minimizing a function, it makes sense to move in the direction opposite to its gradient.

Similarly, we can define the second derivative of the function f, which is generally referred to as the Hessian of f. It is a matrix and its i, j-th entry is given by

$$[\nabla^2 f(x)]_{ij} = \frac{\partial^2 f(x)}{x_i x_j}. \tag{3.2}$$

Using the above definition, for $x, y \in \mathbb{R}^d$ and $A \in \mathbb{R}^{d\times d}$ we obtain

(a) the gradient with respect to x of $x^T y$ is y,
(b) the gradient with respect to x of $x^T x$ is $2x$,
(c) the gradient with respect to x of $x^T Ax$ is $2Ax$,
(d) the gradient with respect to x of Ax is A.

In some cases, you can see that the above gradients are transposed. This is a matter of definition. You should check the wikipedia page https://en.wikipedia.org/wiki/Matrix_calculus which contains a very detailed list of rules.

3.1. *Least squares problem.* In the least squares problem, we are given a target vector $\mathbf{t} \in \mathbb{R}^N$, a design matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$. We would like to find the weights $\mathbf{w}$ that minimizes the objective function given by the least squares problem

$$\underset{\mathbf{w}}{\text{minimize}}\, \mathcal{J}(\mathbf{w}) =: \frac{1}{2}\|\mathbf{t} - \mathbf{X}\mathbf{w}\|_2^2.$$

We know that a minimum occurs at a critical at which the partial derivatives are equal to 0. i.e. $\partial \mathcal{J}(\mathbf{w})/w_j = 0$ for $j = 1, .., D$. This is equivalent to saying the gradient $\nabla \mathcal{J}(\mathbf{w}) = 0$. We can write

$$\mathcal{J}(\mathbf{w}) = \frac{1}{2}\|\mathbf{t}\|_2^2 + \frac{1}{2}\mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} - \mathbf{t}^\top \mathbf{X}\mathbf{w}.$$

Taking derivative with respect to the vector $\mathbf{w}$ and setting it equal to 0, we obtain

$$\nabla \mathcal{J}(\mathbf{w}) = \mathbf{X}^\top \mathbf{X}\mathbf{w} - \mathbf{X}^\top \mathbf{t} = 0.$$

If $\mathbf{X}^\top \mathbf{X}$ is invertible, a solution to above linear system is given by

$$\mathbf{w}^{\text{LS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{t}.$$