This is my preperation notes for teaching in sections during the winter 2018 quarter for course CSE 446. Useful for myself to review the concepts as well.

More Linear Algebra

Definition 1.1 (Dot Product)**.** *(Algebraic definition)* Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two vectors in $\mathbb{R}^n$. Then the dot product (or inner product) between $\boldsymbol{a}$ and $\boldsymbol{b}$ is defined as:

$$\boldsymbol{a} \cdot \boldsymbol{b} = \boldsymbol{a}^T \boldsymbol{b} = \sum_{i=1}^{n} a_i b_i \tag{1}$$

(Geometric definition) The dot product of two Euclidean vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is defined by

$$\boldsymbol{a} \cdot \boldsymbol{b} = |\boldsymbol{a}||\boldsymbol{b}| cos(\theta_{\boldsymbol{a},\boldsymbol{b}}) \tag{2}$$

Also, The dot product $\boldsymbol{w} \cdot \boldsymbol{x} = b$ is a hyperplane, where $\boldsymbol{w}$ is normal to it.

Definition 1.2 (Projection)**.** Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two vectors in $\mathbb{R}^n$. The projection of $\boldsymbol{b}$ onto $\boldsymbol{a}$ is defined

$$proj_{\boldsymbol{a}} \boldsymbol{b} = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{|\boldsymbol{a}|} \frac{\boldsymbol{a}}{|\boldsymbol{a}|} = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{|\boldsymbol{a}|^2} \boldsymbol{a} \tag{3}$$

The projection of a 2-D vector $\boldsymbol{x}$ onto a 1-D line identified by unit vector $\boldsymbol{u}$ is $(\boldsymbol{x} \cdot \boldsymbol{u})\boldsymbol{u}$. To project a N-D vector $\boldsymbol{x}$ down to K-D $\hat{\boldsymbol{x}}$, we have $\hat{\boldsymbol{x}} = \sum_{i=1}^{K} (\boldsymbol{x} \cdot \boldsymbol{u}_i)\boldsymbol{u}_i$.

Definition 1.3 (Outer Product)**.** Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two vectors in $\mathbb{R}^n$. Then the outer product (or tensor product) between $\boldsymbol{a}$ and $\boldsymbol{b}$ is defined such that $(\boldsymbol{a}\boldsymbol{b}^T)_{ij} = a_i b_j$:

$$\boldsymbol{a}\boldsymbol{b}^T = \begin{bmatrix} a_1 b_1 & a_1 b_2 & \cdots & a_1 b_n \\ a_2 b_1 & a_2 b_2 & \cdots & a_2 b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_n b_1 & a_n b_2 & \cdots & a_n b_n \end{bmatrix} \tag{4}$$

Matrix Multiplication

Let $\boldsymbol{A} \in M_{n \times p}(\mathbb{R})$ and $\boldsymbol{B} \in M_{p \times m}(\mathbb{R})$, then $\boldsymbol{AB} \in M_{n \times m}(\mathbb{R})$. Matrix multiplication $\boldsymbol{AB}$ can be interpreted in two ways.

1) When we consider row vectors of $\boldsymbol{A}$ and column vectors of $\boldsymbol{B}$, the multiplication $\boldsymbol{AB}$ can be viewed as

$$\boldsymbol{AB} = \begin{bmatrix} \boldsymbol{A}\boldsymbol{b}_{|1} & \boldsymbol{A}\boldsymbol{b}_{|2} & \cdots & \boldsymbol{A}\boldsymbol{b}_{|m} \end{bmatrix} \tag{5}$$

where $\boldsymbol{B} = \begin{bmatrix} \boldsymbol{b}_{|1} & \boldsymbol{b}_{|2} & \cdots & \boldsymbol{b}_{|m} \end{bmatrix}$. We know $(\boldsymbol{A}\boldsymbol{b}_{|k})_i = \underline{\boldsymbol{a}}_i^T \boldsymbol{b}_{|k}$. Therefore, $(\boldsymbol{AB})_{ij} = \underline{\boldsymbol{a}}_i^T \boldsymbol{b}_{|j}$.

2) When we consider column vectors of $\boldsymbol{A}$ and row vectors of $\boldsymbol{B}$, the multiplication $\boldsymbol{AB}$ can be viewed as

$$\boldsymbol{AB} = \sum_{i=1}^{p} \boldsymbol{a}_{|i} \underline{\boldsymbol{b}}_i^T \tag{6}$$

where $\boldsymbol{a}_{|i} \underline{\boldsymbol{b}}_i^T$ is the *outer product* with output dimension of $n \times m$.

Definition 1.4 (Orthogonal Matrix)**.** An *orthogonal matrix* $\boldsymbol{Q}$ is a square matrix with real entries whose columns and rows are orthogonal unit vectors (i.e., *orthonormal* vectors), i.e.

$$\boldsymbol{Q}^T \boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^T = \boldsymbol{I} \tag{7}$$

Therefore, we have $\boldsymbol{Q}^T = \boldsymbol{Q}^{-1}$. To fully understand why Equation 7 holds, we need to know that for two orthogonal vectors $\boldsymbol{u_1}$ and $\boldsymbol{u_2}$, $\boldsymbol{u_1}^T \boldsymbol{u_2} = 0$. And $\boldsymbol{u_1}^T \boldsymbol{u_1} = |\boldsymbol{u_1}|^2 = 1$. Therefore, in the resulting matrix, all entries are 0 except for ones along the diagonal.

Probability

Parts of this section reference [1].

Event space We define a *space* Ω to be the set of all possible outcomes. An **event space** $\mathcal{S}$ is the set of measurable **events** α such that $\alpha \in \mathcal{S}$ and $\alpha \subseteq \Omega$ to which we are willing to assign probabilities. For example, if we roll a dice, then $\Omega = \{1, 2, 3, 4, 5, 6\}$. A possible event could be $\{1\}$ (we rolled one), $\{1, 3, 5\}$ (we rolled odd), etc. We say an event α *happened* if we observed an outcome $r \in \alpha$. The event space $\mathcal{S}$ is closed under union ($\alpha \in \mathcal{S} \wedge \beta \in \mathcal{S} \rightarrow \alpha \cup \beta \in \mathcal{S}$) and complementation ($\alpha \in \mathcal{S} \rightarrow \Omega - \alpha \subseteq \mathcal{S}$).

Probability distribution Given $(\Omega, \mathcal{S})$, a **probability distribution** $\mathbb{P} : \mathcal{S} \rightarrow \mathbb{R}$ is a mapping from events to real values, such that (1) for all $\alpha \in \mathcal{S}$, $\mathbb{P}(\alpha) \geq 0$, (2) $\mathbb{P}(\Omega) = 1$, and (3) for $\beta \in \mathcal{S}$, $\mathbb{P}(\alpha \cup \beta) = \mathbb{P}(\alpha) + \mathbb{P}(\beta) - \mathbb{P}(\alpha \cap \beta)$.

Random variable A random varialbe $X : \Omega \rightarrow \mathbb{R}$ associates each outcome in Ω with a value. We use $val(X)$ to denote the set of possible values that X can take. Random variables can be discrete or continuous. We primarily consider discrete ones. For simplicity, if x, y are generic values for random variables X and Y, then we write $\mathbb{P}(X = x, Y = y)$ as $\mathbb{P}(X, Y)$. For a specific value x, we write $\mathbb{P}(X = x)$ as $\mathbb{P}(x)$.

Marginal distribution The marginal distribution over random variable X is $\mathbb{P}(X)$.

Joint distribution The joint distribution over random variables $X_1, \cdots X_n$ is $\mathbb{P}(X_1, \cdots, X_n)$ satisfying $\mathbb{P}(X_1) = \sum_{x_i, \cdots x_n} \mathbb{P}(X_1, x_2, \cdots, x_n)$. Note that 1 is arbitrarily chosen.

Conditional probability For random variables X, Y, $\mathbb{P}(X, Y) = \mathbb{P}(X)\mathbb{P}(Y|X)$, where $\mathbb{P}(Y|X)$ is the probability of Y conditioned on X. For random variables $X_1, \cdots, X_n$, we have $\mathbb{P}(X_1, \cdots, X_n) = \mathbb{P}(X_1)(X_2|X_1)\mathbb{P}(X_3|X_1, X_2) \cdots \mathbb{P}(X_n|X_1, \cdots, X_{n-1})$.

Independence X and Y are independent if $\mathbb{P}(X,Y) = \mathbb{P}(X)\mathbb{P}(Y)$. X and Y are *conditionally independent* given Z if $\mathbb{P}(X|Y,Z) = \mathbb{P}(X|Z)$ or $\mathbb{P}(X,Y|Z) = P(X|Z)P(Y|Z)$.

Bayes's Theorem Because $\mathbb{P}(X,Y) = \mathbb{P}(Y)\mathbb{P}(X|Y) = \mathbb{P}(X)P(Y|X)$, we can write

$$\mathbb{P}(X|Y) = \frac{\mathbb{P}(X)\mathbb{P}(Y|X)}{\mathbb{P}(Y)} \tag{8}$$

where $\mathbb{P}(X)$ is called the prior, and $\mathbb{P}(X|Y)$ is called the posterior.

Expectation For discrete random variable X, the expectation of X under distribution $\mathbb{P}$ is defined as

$$\boldsymbol{E}_{\mathbb{P}}[X] = \sum_x x\mathbb{P}(x) \tag{9}$$

We can also write it as $\boldsymbol{E}_X$. Note that we can show $\boldsymbol{E}[f(X)] = \sum_x f(x)\mathbb{P}(x)$. When the subscript is not present, $\boldsymbol{E}[f(X)] = \boldsymbol{E}_X[f(X)]$, and $\boldsymbol{E}[f(X,Y)] = \boldsymbol{E}_{X,Y}[f(X,Y)] = \sum_x \sum_y f(x,y)\mathbb{P}(x,y)$, where X,Y means $\mathbb{P}(X,Y)$, the joint probabilty[1].

Properties of expectation:

- $\boldsymbol{E}_{\mathbb{P}}[aX+b] = a\boldsymbol{E}_{\mathbb{P}}[X] + b$
- $\boldsymbol{E}_{\mathbb{P}}[X+Y] = \boldsymbol{E}_{\mathbb{P}}[X] + \boldsymbol{E}_{\mathbb{P}}[Y]$
- If X and Y are independent, $\boldsymbol{E}_{\mathbb{P}}[XY] = \boldsymbol{E}_{\mathbb{P}}[X]\boldsymbol{E}_{\mathbb{P}}[Y]$
- For a constant value c, $\boldsymbol{E}[c] = c$.

Conditional expectation We define $\boldsymbol{E}_{\mathbb{P}}[X|y] = \sum_x x\mathbb{P}(x|y)$ as the conditional expectation (expectation of X given evidence $Y = y$).

Variance The variance of variable X is

$$\boldsymbol{Var}_{\mathbb{P}}[X] = \boldsymbol{E}_{\mathbb{P}}\Big[(X - \boldsymbol{E}_{\mathbb{P}}[X])^2\Big] = \boldsymbol{E}_{\mathbb{P}}[X^2] - (\boldsymbol{E}_{\mathbb{P}}[X])^2 \tag{10}$$

Properties of variance:

- $\boldsymbol{Var}[aX+b] = a^2\boldsymbol{Var}[X]$
- If X and Y are independent, then $\boldsymbol{Var}_{\mathbb{P}}[X+Y] = \boldsymbol{Var}_{\mathbb{P}}[X] + \boldsymbol{Var}_{\mathbb{P}}[Y]$

Covariance matrix Suppose $X = [X_1, \cdots, X_n]$. Then we define the covariance matrix Σ of X as

$$\Sigma = \boldsymbol{E}[(X - \boldsymbol{E}[X])(X - \boldsymbol{E}[X])^T] \tag{11}$$

if $\mu_i = \boldsymbol{E}[X_i]$, then each entry $\Sigma_{ij} = \text{cov}(X_i, X_j) = \boldsymbol{E}[(X_i - \mu_i)(X_j - \mu_j)] = \boldsymbol{E}[X_iX_j] - \mu_i\mu_j$.

[1] See `http://www2.econ.osaka-u.ac.jp/~tanizaki/class/2012/econome1/05.pdf`.

Bayesian Optimal Classifier

Suppose $\mathcal{D}$ is some distribution of samples (x, y), where $x \in \mathbb{R}^d$ and $y \in val(Y)$ and Y is a discrete random variable. This means $\mathcal{D}(x, y)$ outputs the probability of (x, y) to exist in the world. A classifier $f(x)$ outputs the category (or class) y given input x. The **Bayesian optimal classifier** is one defined as

$$f^*(x) = \arg\max_y \mathcal{D}(x, y) \tag{12}$$

Theorem 3.1. *Bayesian optimal classifier achieve minimal error among all classifiers.*

Proof. Assume there exists another deterministic classifier f' that produces lower error than f^*. Then, for some input x, we have $f'(x) \neq f(x)$. Suppose in the real world, input x can map to classes $\mathcal{Y} = \{y_1, \cdots, y_n\}$. Suppose $f'(x) = y_p \in \mathcal{Y}$. We have:

- Probability of x to occur is $\mathcal{D}(x) = \sum_{y \in \mathcal{Y}} \mathcal{D}(x, y)$.
- Probability of (x, y_p) to be observed is $\mathcal{D}(x, y_p) = \mathcal{D}(x, f'(x))$.
- Probability of (x, y_q) where $y_q \in \mathcal{Y} \setminus \{y_p\}$ to be observed is $\mathcal{D}(x) - \mathcal{D}(x, f'(x))$. This is the probability that f' made a mistake.

Similarly, the probability that f^* made a mistake is $\mathcal{D}(x) - \mathcal{D}(x, f^*(x))$. By definition of Bayesian optimal classifier,

$$\mathcal{D}(x, f^*(x)) = \max_y \mathcal{D}(x, y) \geq \mathcal{D}(x, f'(x)) \tag{13}$$

Therefore,

$$\mathcal{D}(x) - \mathcal{D}(x, f^*(x)) \leq \mathcal{D}(x) - \mathcal{D}(x, f'(x)) \tag{14}$$

Thus, f^* makes fewer mistakes. When f' and f^* only disagree on x, it is not possible for $f'(x)$ being correct while $f^*(x)$ being wrong. Therefore, f^* is optimal. □

Perceptron

Perceptron is one of the simplest (linear) binary classifiers. Suppose we observe data points (x_i, y_i) for $i = 1, \cdots, N$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$. We hope to train the weights $w \in \mathbb{R}^d$ and bias b in the following classification function:

$$\hat{y} = f(x) = \text{sign}(w \cdot x + b) \tag{15}$$

To minimize a loss function $L(w) = \frac{1}{N} \sum_{i=1}^{N} \ell(y_i, \hat{y})$. The perceptron algorithm is an iterative process, either offline or online. See Algorithm 1 and 2.

Perceptron is a linear classifier, which means the decision boundary is a linear combination of features (i.e. a hyperplane). Therefore, if the data is not linearly separable (cannot

Algorithm 1: Perceptron-Online(T)

$w^{(1)} \leftarrow \mathbf{0}$; $b^{(1)} \leftarrow 0$;
foreach $t = 1, \cdots, T$ **do**
  $(x_t, y_t) \leftarrow$ new observation at time t;
  $\hat{y}_t \leftarrow f(w^{(t)} \cdot x_t)$;
  if $\hat{y}_t \neq y_t$ **then**
    $w^{(t+1)} \leftarrow w^{(t)} + y_t x_t$;
    $b^{(t+1)} \leftarrow b^{(t)} + y_t$
  end
end

Algorithm 2: Perceptron-Offline($\mathcal{D}, T$)

$w^{(1)} \leftarrow \mathbf{0}$; $b^{(1)} \leftarrow 0$;
foreach $t = 1, \cdots, T$ **do**
  foreach $(x_i, y_i) \in \mathcal{D}$ **do**
    $\hat{y}_i^{(t)} \leftarrow f(w^{(t)} \cdot x_i)$;
    if $\hat{y}_i^{(t)} \neq y_i$ **then**
      $w^{(t+1)} \leftarrow w^{(t)} + y_i x_i$;
      $b^{(t+1)} \leftarrow b^{(t)} + y_i$
    end
  end
end

be separated by a hyperplane), then perceptron will not converge. If the data is indeed linearly separable, then perceptron is guaranteed to converge.

Prove that perceptron is guaranteed to converge.[2] Assume the data is linearly separable. The definition of linear separability tells us that there exists some weights w^* and the decision boundary $w^* \cdot x$ separates the data with a margin $\gamma \geq 0$, i.e.,

$$y_i(w^* \cdot x) \geq \gamma \tag{16}$$

Suppose we have weights $w^{(t+1)}$ at time $t+1$. We are interested to know if inequality $||w^{(t+1)} - w^*||^2 \leq ||w^{(t)} - w^*||^2$ holds. That is, $w^{(t+1)}$ is "closer" to w^*, the weights that can be used to separate the data. Let binary variable $m_i = 1$ if there is a mistake when

[2]Proof learned from Sham Kakade's notes: `https://courses.cs.washington.edu/courses/cse546/16au/slides/notes09.pdf`. There is another proof [Novikoff] that shows a better upper bound for the number of mistakes, but it is not necessary for our problem.

classifying data point i.

$$||w^{(t+1)} - w^*||^2 = ||w^{(t)} + m_i y_i x_i - w^*||^2 \tag{17}$$
$$= ||w^{(t)} - w^*||^2 + 2m_i y_i x_i^T (w^{(t)} - w^*) + m_i^2 y_i^2 ||x_i||^2 \tag{18}$$
$$\leq ||w^{(t)} - w^*||^2 + 2m_i y_i x_i^T (w^{(t)} - w^*) + m_i^2 \tag{19}$$
$$\leq ||w^{(t)} - w^*||^2 - 2m_i + m_i \tag{20}$$
$$\leq ||w^{(t)} - w^*||^2 - m_i \tag{21}$$

(19) to (20) is because, from (16), we have $y_i x_i^T w^* \geq 0$. And because $m_i y_i x_i^T w^{(t)} \leq 0$, we have

$$m_i y_i x_i^T (w^{(t)} - w^*) \leq m_i y_i x_i^T w^{(t)} - m_i \leq -m_i \tag{22}$$

Therefore,

$$m_i \leq ||w^{(t)} - w^*||^2 - ||w^{(t+1)} - w^*||^2 \tag{23}$$

Perceptron is indeed improving at every iteration. Suppose the total number of mistakes at iteration T si $M_T = \sum_{i=1}^{T} m_i$. From the above inequality, we arrive at

$$M_T \leq ||w^{(1)} - w^*||^2 - ||w^{(T+1)} - w^*||^2 \leq ||w^*||^2 \tag{24}$$

The last inequality holds because $w^{(1)} = \mathbf{0}$. Therefore, there is an upper-bounded for the total number of mistakes, which means perceptron is guaranteed to converge.

References

[1] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.