Kyle Kazemini
CS 3190-001
12/11/20

Final Cheat Sheet

Ch 1, 2 $P(A|B) = \frac{P(A \cap B)}{P(B)}$ $\quad E[X] = \sum_{\omega \in \Omega} (\omega \, P[X = \omega])$

$E[X] = \int_{\omega \in \Omega} \omega \, f_X(\omega) \, d\omega$ $\quad Var[X] = E[X^2] - E[X]^2$

CLT: $X_1, X_2, \ldots, X_n$ $\quad X_i \sim f$ $\quad \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
converges to a Normal distribution with
$\mu = E[X_i]$ & variance $\frac{\sigma^2}{n}$

PAC: $P[\, |\bar{X} - E[\bar{X}]| \geq \varepsilon \,] \leq \delta$

Markov: $P[X > \alpha] \leq \frac{E[X]}{\alpha}$

Chebyshev: $P[\, |X - E[X]| \geq \varepsilon \,] \leq \frac{Var[X]}{\varepsilon^2}$

Chernoff-Hoeffding: $P[\, |\bar{X} - E[\bar{X}]| > \varepsilon \,] \leq 2\exp\left(\frac{-2\varepsilon^2 n}{\Delta^2}\right)$

Ch 3 $\|v\| = \|v\|_2 = \sqrt{\langle v, v \rangle}$, $\|v\|_p = \left(\sum_{i=1}^{d} |v_i|^p\right)^{1/p}$ $\quad v \in \mathbb{R}^d$

$\|A\|_2 = \max_{\substack{x \in \mathbb{R}^d \\ x \neq 0}} \frac{\|Ax\|_2}{\|x\|_2} = \max_{\substack{y \in \mathbb{R}^n \\ y \neq 0}} \frac{\|yA\|_2}{\|y\|_2}$

$\|A\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{d} A_{ij}^2}$

Ch 5

$y = \ell(x) = ax + b$.

$\forall x \in \mathbb{R}$, we can predict a value $\hat{y} = \ell(x)$

The line ℓ is our model for this input data.

Measure error:

$$r_i = y_i - \hat{y}_i = y_i - \ell(x_i)$$

$$SSE((X,y),\ell) = \sum_{i=1}^{n} r_i^2$$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - \ell(x_i))^2$$

$\hat{y}_i = M_\alpha(x_i)$ where $\alpha = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T y$

$\hat{y} = M_p(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \dots + \alpha_p x^p$

Soln: $\tilde{X} \in \mathbb{R}^{n \times (p+1)}$

Cross validation: $\alpha = (X_{train}^T X_{train})^{-1} X_{train}^T Y_{train}$

Soln: $\ell : x \to \tilde{x} \in \mathbb{R}^{n \times (d+1)}$

Soln: $\alpha = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T y$

Goal: $\alpha^* = \underset{\alpha \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \|\tilde{X}\alpha - y\|^2$

Ch 6 $f(\alpha) = f(\alpha_1, \alpha_2, \ldots, \alpha_d)$, $u = (u_1, u_2, \ldots, u_d)$

$$\nabla_u f(\alpha) = \lim_{h \to 0} \frac{f(\alpha + hu) - f(\alpha)}{h}$$

$e_1, e_2, \ldots, e_d$ a set of unit vectors

$$\nabla f(\alpha) = \frac{\partial f}{\partial \alpha_1} e_1 + \frac{\partial f}{\partial \alpha_2} e_2 + \ldots + \frac{\partial f}{\partial \alpha_d} e_d$$

$\nabla f : \mathbb{R}^d \to \mathbb{R}^d$, $\nabla_u f(\alpha) = \langle \nabla f(\alpha), u \rangle$

Gradient Descent:

Goal: $\min_{\alpha \in \mathbb{R}^d} f(\alpha)$ and/or $\alpha^* = \operatorname{argmin}_{\alpha \in \mathbb{R}^d} f(\alpha)$

0: Initialize $\alpha^{(0)} = \alpha_{start} \in \mathbb{R}^d$, $k = 0$

1: Repeat $\alpha^{(k+1)} = \alpha^{(k)} - \gamma_k \nabla f(\alpha^{(k)})$, $k+1$

until: $\| \nabla f(\alpha^{(k)}) \| \leq \tau$ or $k = T$

2: return $\alpha^{(k)}$

Here, γ is the learning rate, τ is some tolerance level, T is some number of iterations.

Ch 8 **Clustering**: Input $X = \{x_1, x_2, \ldots, x_n\}$ $X \subset \mathbb{R}^d$

Distance $D: X \times X \to \mathbb{R}^+$ $\quad D(x_1, x_2) = \|x_1 - x_2\|$

Goal: K subsets $\{X_1, X_2, \ldots, X_K\}$ $X_i \subset X$

$$\phi_S(x) = \underset{s_i \in S}{\operatorname{argmin}} \|x - s_i\|$$

$$\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \|x_i - \phi_S(x_i)\|^2$$

Lloyd's algorithm: $\text{cost}(X, S) = \sum_{x \in X} \|\phi_S(x) - x\|^2$

1) Choose k points $S \subset X$
2) $\forall x \in X$, assign x to X_i so $\phi_S(x) = s_i$
3) $\forall s_i \in S$, update $s_i = \frac{1}{|X_i|} \sum_{x \in X_i} x$
4) until S is unchanged.

Mixture of Gaussians covariance matrix:

$$\Sigma_i = \sum_{x \in X_i} (x - \mu_i)(x - \mu_i)^T$$

Loss function:

$$f(\alpha) = \mathcal{L}(g_\alpha, (X, y)) = \sum_{i=1}^{n} f_i(\alpha) \quad \text{where}$$

$$f_i(\alpha) = \ell(z_i = y_i g_\alpha(x_i))$$

Ch 8

Perceptron algorithm:

Initialize $w = y_i x_i$ for any x_i, y_i in (X, Y)

Repeat:

$$\forall (x_i, y_i) \text{ such that } y_i \langle x_i, w \rangle < 0,$$

update

$$w \leftarrow w + y_i x_i$$

until T steps, or there are no more misclassifications.

Return

$$w \leftarrow \frac{w}{\|w\|}$$

Ch 7

Recall:

Data as a matrix:

$A \in \mathbb{R}^{n \times d} \rightarrow$ SVD $\rightarrow$ map to each

$f(a_i) = b_i \in \mathbb{R}^d \rightarrow \mathbb{R}^k$

Projection:

$$\Pi_B(a) = \sum_{j=1}^{k} \langle v_j, a \rangle v_j$$

$$\Pi_F(p) = \Pi_B(a) = \sum_{j=1}^{k} \Pi_{v_j}(p)$$

Where

$$B^* = \underset{B}{\operatorname{argmin}} \; SSE(A, B)$$

Ch 7 $SSE(A,B) = \sum_{a_i \in A} \|a_i - \pi_B(a_i)\|^2$

SVD: $A \in \mathbb{R}^{n \times d}$, $U \in \mathbb{R}^{n \times n}$, $S \in \mathbb{R}^{n \times d}$, $V \in \mathbb{R}^{d \times d}$

$$A = USV^T$$

PCA: k-dimensional subspace B to minimize

$$\|A - \pi_B(A)\|_F^2 = \sum_{a_i \in A} \|a_i - \pi_B(a_i)\|^2$$

Power method: Input $A \in \mathbb{R}^{n \times d}$
$M = A^TA \in \mathbb{R}^{d \times d}$ positive semi-definite

(Positive semi-definite means non-negative eigenvalues)

$v \leftarrow$ random vector in $\mathbb{R}^d$

$v = M^q u$ for $i = 1, \ldots, q$

$u^{(i)} = Mu^{(i-1)}$

return $v_1 = \frac{v}{\|v\|_2}$