Machine Learning 1 - Cheat Sheet

Multivariate Calculus

Index notation

- $[\mathbf{A}\mathbf{v}]_i = \sum_p \mathbf{A}_{ip}\mathbf{v}_p$
- $\mathbf{v}^{\mathrm{T}}\mathbf{A}\mathbf{x} = \sum_p \sum_q \mathbf{v}_p \mathbf{A}_{pq}\mathbf{x}_q$
- $\mathbf{v}^{\mathrm{T}}\mathbf{x} = \sum_p \mathbf{v}_p \mathbf{x}_p$

Multivariate derivatives

- $\frac{\partial}{\partial \mathbf{w}}\mathbf{w}^{\mathrm{T}}\mathbf{v} = \mathbf{v}^{\mathrm{T}}$
- $\frac{\partial}{\partial \mathbf{w}}\mathbf{w}^{\mathrm{T}}\mathbf{A}\mathbf{w} = \mathbf{w}^{\mathrm{T}}(\mathbf{A} + \mathbf{A}^{\mathrm{T}})$
- $\frac{\partial}{\partial \mathbf{w}}\mathbf{A}\mathbf{w} = \mathbf{A}$

Useful functions

- Kronecker delta: $\delta_{ik} = \begin{cases} 1 & \text{if } i = k \\ 0 & \text{otherwise} \end{cases}$
- Indicator function: $\mathbf{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$

Conventions

- Vectors are columns ($\mathbf{x} \in \mathbb{R}^{n \times 1}$)
- If $f : \mathbb{R}^n \longrightarrow \mathbb{R}^m$, then $\frac{df}{d\mathbf{x}} \in \mathbb{R}^{m \times n}$

Constrained optimization

Equality constraint

$$\max_{\mathbf{x}} f(\mathbf{x}) \text{ subject to } g(\mathbf{x}) = 0$$

- Lagrangian: $L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})$.

Inequality constraint

$$\max_{\mathbf{x}} f(\mathbf{x}) \text{ subject to } g(\mathbf{x}) \geq 0$$

- Lagrangian: $L(\mathbf{x}, \mu) = f(\mathbf{x}) + \mu g(\mathbf{x})$.
- Solve $\max_{\mathbf{x}} \min_{\mu} L(\mathbf{x}, \mu)$ subject to KKT cond.:

$$g(\mathbf{x}) \geq 0, \quad \mu \geq 0, \quad \mu g(\mathbf{x}) = 0.$$

Probability & Statistics

Probability

- Sum rule: $P(X) = \sum_Y P(X, Y)$ (disc.)
- Product rule: $P(X, Y) = P(X \mid Y)P(Y)$

Moments

- $\mathbb{E}[f(X)] = \int_x f(x)p(x)dx$ (cont.)
- $\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
- $\mathrm{Cov}[X, Y] = \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big]$
  $= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

Distributions will be provided if needed.

Regression

Linear Regression with Basis Functions

- Model: $t = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}) + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \beta^{-1})$
- Least sq. sol.: $\hat{\mathbf{w}} = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}$
- Reg. least sq. sol.: $\hat{\mathbf{w}} = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi} + \lambda \mathbf{I}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}$

where

- Design matrix: $\mathbf{\Phi} = \left(\boldsymbol{\phi}(\mathbf{x}_1), \boldsymbol{\phi}(\mathbf{x}_2), \ldots\right)^{\mathrm{T}}$

Unsupervised methods

PCA

- Eigen-decompostion: $\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\top}$.
- Projection: $\mathbf{z} = \mathbf{U}_M^{\top}(\mathbf{x} - \bar{\mathbf{x}})$.
- Whitened projection: $\mathbf{z} = \mathbf{\Lambda}_M^{-1/2}\mathbf{U}_M^{\top}(\mathbf{x} - \bar{\mathbf{x}})$.

Probabilistic PCA

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

Mixture of experts

$$p(\mathbf{x}) = \sum_k p(\mathbf{x} \mid z_k = 1)p(z_k = 1)$$

- Responsibility: $\gamma(z_k) := p(z_k = 1 \mid \mathbf{x})$

Classification

Logistic Regression

- Sigmoid function: $\sigma(z) = \frac{1}{1+e^{-z}}$
- Softmax function: $\varsigma(\mathbf{z})_i = \frac{\exp z_i}{\sum_{j=1}^{n} \exp z_j}$

Cross-entropy loss

$$E = -\sum_{n=1}^{N}\sum_{k=1}^{K} y_{nk}\log(\hat{y}_{nk})$$

Soft margin classifier

$$\operatorname*{arg\,min}_{\mathbf{w}, b, \xi_n} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n$$

$$\text{subject to} \quad \begin{aligned} & t_n y(\mathbf{x}_n) \geq 1 - \xi_n, && \forall n \in \{1, ..., N\}, \\ & \xi_n \geq 0, && \forall n \in \{1, ..., N\}. \end{aligned}$$

Kernel methods

Kernels

- $\mathbf{K}$ ($K_{nm} = k(\mathbf{x}_n, \mathbf{x}_n)$) must be symmetric positive semi definite for k to be a valid kernel.
- Given valid kernels $k_1(\mathbf{x}, \mathbf{x}')$ and $k_2(\mathbf{x}, \mathbf{x}')$, the following new kernels will also be valid:

  $c\,k_1(\mathbf{x}, \mathbf{x}')$, $f(\mathbf{x})\,k_1(\mathbf{x}, \mathbf{x}')\,f(\mathbf{x}')$, $q(k_1(\mathbf{x}, \mathbf{x}'))$,
  $\exp(k_1(\mathbf{x}, \mathbf{x}'))$, $k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}')$, $k_1(\mathbf{x}, \mathbf{x}')\,k_2(\mathbf{x}, \mathbf{x}')$,
  $k_3(\boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}'))$, $\mathbf{x}^{\top}A\mathbf{x}'$, $k_a(\mathbf{x}_a, \mathbf{x}'_a) + k_b(\mathbf{x}_b, \mathbf{x}'_b)$,
  $k_a(\mathbf{x}_a, \mathbf{x}'_a)\,k_b(\mathbf{x}_b, \mathbf{x}'_b)$.

  where $c > 0$ is a constant, $f(\cdot)$ is any function, $q(\cdot)$ is a polynomial with nonnegative coefficients, $\boldsymbol{\phi}(\mathbf{x}) : \mathbf{x} \to \mathbb{R}^M$, $k_3(\cdot, \cdot)$ is a valid kernel in $\mathbb{R}^M$, and A is symmetric positive semidefinite. For $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_b)$, k_a and k_b are valid kernel functions over their respective spaces.

Gaussian processes

$$f(\cdot) \sim GP(m(\cdot), k(\cdot, \cdot))$$