MA4270 Data Modelling and Computation

Final Examination Helpsheet

AY2023/24 Semester 2 · Prepared by Tian Xiao *@snoidetx*

1 Perceptron

Classification Problems: To learn a classifier $f_{\boldsymbol{\theta}}$ that classifies labels.

- Dataset: $\mathcal{D} = \{(\mathbf{x}_t, y_t)\}_{t=1}^n$ where $\mathbf{x}_t \in \mathbb{R}^d$ and $y_t \in \{-1, +1\}$.
- Classifier: $f_{\boldsymbol{\theta}} : \mathbb{R}^d \to \{-1, +1\}$.
  - ▷ Linear classifier: $f_{\boldsymbol{\theta}} = \text{sign}\left(\boldsymbol{\theta}^\top \mathbf{x}\right)$.
- Training error: $\hat{E}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^n \text{Loss}(y_t, f_{\boldsymbol{\theta}}(\mathbf{x}_t))$.
  - ▷ 0-1 loss: $\text{Loss}(y, \hat{y}) = \ell(y, \hat{y}) = \mathbf{1}\{\hat{y} \neq y\} = \begin{cases} 1 & \hat{y} \neq y \\ 0 & \text{otherwise} \end{cases}$.
  - ▷ Linear separable: $\exists \boldsymbol{\theta}\ [\hat{E}(\boldsymbol{\theta}) = 0]$.

The Perceptron Algorithm:

① Initialize $\boldsymbol{\theta}^{(0)}$ to some value (e.g., $\mathbf{0}$), and initialize index k to 0.
② Repeatedly perform the following:
- ▷ Select the next example $(\mathbf{x}_t, y_t)$ from the training set and check whether $\boldsymbol{\theta}^{(k)}$ classifies it correctly.
- ▷ If it is incorrect (i.e., $y_t\left(\boldsymbol{\theta}^{(k)}\right)^\top \mathbf{x}_t < 0$), set $\boldsymbol{\theta}^{(k+1)} \leftarrow \boldsymbol{\theta}^{(k)} + y_t\mathbf{x}_t$ and increment $k \leftarrow k+1$.

- Assumptions:
  ① Inputs are bounded: $\exists R \in (0, \infty)\ \forall \mathbf{x}_t \in \mathcal{D}\ [\|\mathbf{x}_t\| \leq R]$.
  ② Linearly separable: $\exists \boldsymbol{\theta}^*\ \exists \gamma > 0\ [\min_t y_t\left(\boldsymbol{\theta}^*\right)^\top \mathbf{x}_t \geq \gamma]$.
- **Convergence.** Under the initial vector $\boldsymbol{\theta}^{(0)} = \mathbf{0}$, for any dataset $\mathcal{D}$ satisfying the above assumptions, the perceptron algorithm produces a vector $\boldsymbol{\theta}^{(k)}$ classifying every example correctly after at most $k_{\max} = \frac{R^2\|\boldsymbol{\theta}^*\|^2}{\gamma^2}$ mistakes (and hence update steps).

Proof. Let $R = \max\|\mathbf{x}_t\|$, $\gamma = \min y_t(\boldsymbol{\theta}^*)^\top \mathbf{x}_t$ for $t = 1, 2, \cdots, n$.
① $(\boldsymbol{\theta}^*)^\top\boldsymbol{\theta}^{(k)} = (\boldsymbol{\theta}^*)^\top\left(\boldsymbol{\theta}^{(k-1)} + y_t\mathbf{x}_t\right) \geq (\boldsymbol{\theta}^*)^\top\boldsymbol{\theta}^{(k-1)} + \gamma$. So $(\boldsymbol{\theta}^*)^\top\boldsymbol{\theta}^{(k)} \geq k\gamma$.
② $\|\boldsymbol{\theta}^{(k)}\|^2 = \|\boldsymbol{\theta}^{(k-1)}\|^2 + 2\langle\boldsymbol{\theta}^{(k-1)}, y_t\mathbf{x}_t\rangle + \|\mathbf{x}_t\|^2 \leq \|\boldsymbol{\theta}^{(k-1)}\|^2 + \|\mathbf{x}_t\|^2$. So $\|\boldsymbol{\theta}^{(k)}\|^2 \leq kR^2$.
③ By Cauchy-Schwarz inequality $\langle\mathbf{v}, \mathbf{w}\rangle \leq \|\mathbf{v}\|\cdot\|\mathbf{w}\|$, we have $1 \geq \frac{\langle\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^*\rangle}{\|\boldsymbol{\theta}^{(k)}\|\cdot\|\boldsymbol{\theta}^*\|} \geq \frac{k\gamma}{\sqrt{kR^2}\|\boldsymbol{\theta}^*\|}$, hence $k \leq \frac{R^2\|\boldsymbol{\theta}^*\|^2}{\gamma^2}$.

- Margin: Let $\gamma = \min_{t=1,2,\cdots,n} y_t\boldsymbol{\theta}^\top\mathbf{x}_t$. The quantity $\gamma_{\text{geom}} = \frac{\gamma}{\|\boldsymbol{\theta}\|}$ is the smallest distance from any example $\mathbf{x}_t$ to the decision boundary specified by $\boldsymbol{\theta}$.

2 Support Vector Machine (SVM)

Maximum Margin Classifier: $\min_{\boldsymbol{\theta}} \frac{1}{2}\|\boldsymbol{\theta}\|^2\ \text{s.t.}\ \forall t\ \left[y_t\boldsymbol{\theta}^\top\mathbf{x}_t \geq 1\right]$ (unique).

- SVM with offset: $\min_{\boldsymbol{\theta},\theta_0} \frac{1}{2}\|\boldsymbol{\theta}\|^2$ s.t. $\forall t\ \left[y_t\left(\boldsymbol{\theta}^\top\mathbf{x}_t + \theta_0\right) \geq 1\right]$.
  - ▷ Support vectors: On margin $(y_t\left(\boldsymbol{\theta}^\top\mathbf{x}_t + \theta_0\right) = 1)$.
- Soft-margin SVM: $\min_{\boldsymbol{\theta},\theta_0,\zeta} \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{t=1}^n \zeta_t$ s.t. $\forall t\ \left[y_t\left(\boldsymbol{\theta}^\top\mathbf{x}_t + \theta_0\right) \geq 1 - \zeta_t\right]$.
  - ▷ $\boldsymbol{\zeta} = (\zeta_1, \zeta_2, \cdots, \zeta_n) \geq \mathbf{0}$ is called *slack variables*.
  - ▷ Support vectors: On margin/within margin/misclassified.
- Hinge-loss formulation: $\min_{\boldsymbol{\theta},\theta_0} \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{t=1}^n \left[1 - y_t\left(\boldsymbol{\theta}^\top\mathbf{x}_t + \theta_0\right)\right]_+$.
  - ▷ Hinge loss: $z \to [1-z]_+ = \max\{0, 1-z\}$.
  - ▷ Interpretation: Total hinge loss with regularization term $\frac{1}{2}\|\boldsymbol{\theta}\|^2$.

3 Logistic Regression

Logistic Likelihood Model: $\Pr(y \mid \mathbf{x}) = \frac{1}{1+\exp(-y(\boldsymbol{\theta}^\top\mathbf{x}+\theta_0))}$.

- $g(z) = \frac{1}{1+e^{-z}} \in (0, 1)$ assigns *likelihood* to points.
  - ▷ Scaling the dataset by $c > 1$ pushes prediction closer to 0 or 1.
  - ▷ Linear classifier chooses the label that is more likely under the logistic model.
  - ▷ Log-odds $\log\frac{\Pr(y=1|\mathbf{x})}{\Pr(y=-1|\mathbf{x})}$ is a linear function $\langle\boldsymbol{\theta}, \mathbf{x}\rangle + \theta_0$ of inputs.
- Maximum likelihood estimate (MLE) of parameters:

$$\begin{aligned}
\left(\hat{\boldsymbol{\theta}}, \hat{\theta}_0\right) &= \arg\max_{\boldsymbol{\theta},\theta_0} \prod_{t=1}^n \Pr(y_t \mid \mathbf{x}_t; \boldsymbol{\theta}, \theta_0) \quad \text{(likelihood)} \\
&= \arg\max_{\boldsymbol{\theta},\theta_0} \prod_{t=1}^n \frac{1}{1+\exp(-y_t(\boldsymbol{\theta}^\top\mathbf{x}_t + \theta_0))} \quad \text{(likelihood)} \\
&= \arg\max_{\boldsymbol{\theta},\theta_0} \sum_{t=1}^n \log\frac{1}{1+\exp(-y_t(\boldsymbol{\theta}^\top\mathbf{x}_t + \theta_0))} \quad \text{(log-likelihood)} \\
&= \arg\min_{\boldsymbol{\theta},\theta_0} \sum_{t=1}^n \log\left(1+\exp\left(-y_t\left(\boldsymbol{\theta}^\top\mathbf{x}_t + \theta_0\right)\right)\right).
\end{aligned}$$

- Regularization: $\min_{\boldsymbol{\theta},\theta_0} \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{t=1}^n \log\left(1+\exp\left(-y_t\left(\boldsymbol{\theta}^\top\mathbf{x}_t + \theta_0\right)\right)\right)$.
  - ▷ Logistic loss: $z \to \log\left(1+e^{-z}\right)$.
  - ▷ Interpretation: Total logistic loss with regularization term $\frac{1}{2}\|\boldsymbol{\theta}\|^2$.
- Softmax function (multiclass): $\Pr(y = c \mid \mathbf{x}) = \frac{\exp\left(\boldsymbol{\theta}_c^\top\mathbf{x}+\theta_{0,c}\right)}{\sum_{c'=1}^M \exp\left(\boldsymbol{\theta}_{c'}^\top\mathbf{x}+\theta_{0,c'}\right)}$.
  - ▷ When $M = 2$, we recover logistic model by setting $(\boldsymbol{\theta}_c, \theta_{0,c}) = (\mathbf{0}, 0)$ for one of the two classes.

4 Linear Regression

Linear Predictor: $\hat{y} = \boldsymbol{\theta}^\top\mathbf{x} + \theta_0$.

- Matrix form: $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\Theta}$, where $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1^\top & 1 \\ \vdots & \vdots \\ \mathbf{x}_n^\top & 1\end{bmatrix}$ and $\boldsymbol{\Theta} = \begin{bmatrix}\boldsymbol{\theta} \\ \theta_0\end{bmatrix}$.
- Least squares estimate (LSE): $\hat{\boldsymbol{\Theta}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$.
  - ▷ Unique solution if $\mathbf{X}^\top\mathbf{X}$ is invertible.
- Gaussian model: $y_t = (\boldsymbol{\theta}^*)^\top\mathbf{x}_t + \theta_0^* + z_t$, where $z_t \sim \mathcal{N}(0, \sigma^2)$.
  - ▷ Gaussian PDF: $\mathcal{N}(z; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right)$.
  - ▷ $\Pr(y \mid \mathbf{x}) = \mathcal{N}(y; (\boldsymbol{\theta}^*)^\top\mathbf{x} + \theta_0^*, \sigma^2)$.
  - ▷ Log-likelihood:
    $$\log\prod_{t=1}^n \Pr(y_t \mid \mathbf{x}_t) = \text{const.} - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^n\left(y_t - \boldsymbol{\theta}^\top\mathbf{x}_t - \theta_0\right)^2.$$
  - ▷ MLE of $\boldsymbol{\theta}$ and θ_0: $\left(\hat{\boldsymbol{\theta}}, \hat{\theta}_0\right) = \arg\min_{\boldsymbol{\theta},\theta_0}\sum_{t=1}^n\left(y_t - \boldsymbol{\theta}^\top\mathbf{x}_t - \theta_0\right)^2$.
    - ∗ σ^2 is assumed to be known.
    - ∗ MLE of σ^2: $\hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^n\left(y_t - \hat{\boldsymbol{\theta}}^\top\mathbf{x}_t - \hat{\theta}_0\right)^2$.
- Gaussian model in matrix form: $\mathbf{y} = \mathbf{X}\boldsymbol{\Theta}^* + \mathbf{z}$.
  - ▷ LSE: $\hat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}^* + \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{z}$.
    - ∗ No bias: $\mathbb{E}[\hat{\boldsymbol{\Theta}}] = \boldsymbol{\Theta}^*$.
    - ∗ Covariance: $\text{Cov}[\hat{\boldsymbol{\Theta}}] = \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$.
- Ridge regression: $\left(\hat{\boldsymbol{\theta}}, \hat{\theta}_0\right) = \arg\min_{\boldsymbol{\theta},\theta_0}\sum_{t=1}^n\left(y_t - \boldsymbol{\theta}^\top\mathbf{x}_t - \theta_0\right)^2 + \lambda\sum_{j=1}^d\theta_j^2$.
  - ▷ Closed-form solution (w/o offset): $\hat{\boldsymbol{\theta}} = \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{y}$.
    - ∗ $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$ is always invertible when $\lambda > 0$.
  - ▷ Assuming no offset θ_0:
    - ∗ Bias: $\mathbb{E}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}^* = -\lambda\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\boldsymbol{\theta}^*$.
    - ∗ Covariance: $\sigma^2\left(\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)^{-1} - \lambda\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)^{-2}\right)$.

Bias-Variance Tradeoff: Decomposition of MSE:

$$\mathbb{E}[\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|^2] = \underbrace{\|\mathbb{E}[\hat{\boldsymbol{\Theta}}] - \boldsymbol{\Theta}^*\|^2}_{\text{bias}} + \underbrace{\mathbb{E}[\|\hat{\boldsymbol{\Theta}} - \mathbb{E}[\hat{\boldsymbol{\Theta}}]\|^2]}_{\text{variance}}.$$

Proof. Let $\boldsymbol{\mu} = \mathbb{E}[\hat{\boldsymbol{\Theta}}]$.
① bias $= \|\boldsymbol{\mu}\|^2 - 2\langle\boldsymbol{\mu}, \boldsymbol{\Theta}^*\rangle + \|\boldsymbol{\Theta}^*\|^2$.
② variance $= \mathbb{E}[\|\hat{\boldsymbol{\Theta}}\|^2 - 2\langle\hat{\boldsymbol{\Theta}}, \boldsymbol{\mu}\rangle + \|\boldsymbol{\mu}\|^2] = \mathbb{E}[\|\hat{\boldsymbol{\Theta}}\|^2] - 2\langle\mathbb{E}[\hat{\boldsymbol{\theta}}], \boldsymbol{\mu}\rangle + \|\boldsymbol{\mu}\|^2 = \mathbb{E}[\|\hat{\boldsymbol{\Theta}}\|^2] - \|\boldsymbol{\mu}\|^2$.
③ bias + variance $= \mathbb{E}[\|\hat{\boldsymbol{\Theta}}\|^2] - 2\langle\boldsymbol{\mu}, \boldsymbol{\Theta}^*\rangle + \|\boldsymbol{\Theta}^*\|^2 = \text{LHS}$.

5 Kernel Method

Kernel: A measure of similarity.

- Kernel matrix: A function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is said to be a *positive semidefinite* (PSD) kernel if
  ① it is symmetric, i.e., $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}', \mathbf{x})$;
  ② the following *kernel matrix* is always PSD:
  $$\mathbf{K} = \begin{bmatrix}k(\mathbf{x}_1, \mathbf{x}_1) & \cdots & k(\mathbf{x}_1, \mathbf{x}_m) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_m, \mathbf{x}_1) & \cdots & k(\mathbf{x}_m, \mathbf{x}_m)\end{bmatrix}.$$
  - ▷ Polynomial kernel: $k(\mathbf{x}, \mathbf{x}') = (1 + \langle\mathbf{x}, \mathbf{x}'\rangle)^p$.
  - ▷ RBF kernel: $k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2}\|\mathbf{x} - \mathbf{x}'\|^2\right)$.
- k is a PSD kernel iff it equals an inner product $\langle\phi(\mathbf{x}), \phi(\mathbf{x}')\rangle$ for some (possibly infinite dimensional) mapping ϕ.
- Construction: If k_1, k_2 are kernels, then the following are kernels:
  ① $k(\mathbf{x}, \mathbf{x}') = f(\mathbf{x})k_1(\mathbf{x}, \mathbf{x}')f(\mathbf{x}')$ for some function f;
  ② $k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}')$;
  ③ $k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}')$.

Kernel Trick: Replace $\langle\mathbf{x}, \mathbf{x}'\rangle$ by $k(\mathbf{x}, \mathbf{x}')$.

- Possible when algorithm depends on only inputs' inner products.
- Dual → Kernel SVM: $\alpha > 0$ support vectors; $\alpha = C$ violation.

(P) $\min_{\boldsymbol{\theta},\theta_0,\zeta}\ \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^n\zeta_t$ s.t. $y_t(\boldsymbol{\theta}^\top\mathbf{x}_t + \theta_0) \geq 1 - \zeta_t$; $\zeta_t \geq 0, \forall t$.

(D) $\max_{\boldsymbol{\alpha}}\ \sum_{t=1}^n\alpha_t - \frac{1}{2}\sum_{s=1}^n\sum_{t=1}^n\alpha_s\alpha_t y_s y_t \underbrace{\mathbf{x}_s^\top\mathbf{x}_t}_{k(\mathbf{x}_s, \mathbf{x}_t)}$ s.t. $\alpha_t \in [0, C], \forall t$; $\sum_{t=1}^n\alpha_t y_t = 0$.

6 Gradient-Based Optimization

Gradient Descent: W.r.t. $f(\mathbf{x})$, $\mathbf{x}_{\text{next}} = \mathbf{x} - \eta\cdot\nabla f(\mathbf{x})$.

- Stochastic gradient descent (SGD): $\mathbf{x}_{\text{next}} = \mathbf{x} - \eta \cdot \nabla f_i(\mathbf{x})$.
- Mini-batch SGD: $\mathbf{x}_{\text{next}} = \mathbf{x} - \eta \cdot \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \nabla f_i(\mathbf{x})$.

Subgradient-Based Optimization: Non-differentiable convex functions.

- Subgradient: $\partial f(\mathbf{x}) = \{\mathbf{g} \in \mathbb{R}^d : f(\mathbf{x}') \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{x}' - \mathbf{x} \rangle, \forall \mathbf{x}'\}$.
  - ▷ $f(x) = |x|$: $\partial = \{1\}$ at $x > 0$; $\{-1\}$ at $x < 0$; $[-1, 1]$ at $x = 0$.
- Subgradient method: $\mathbf{x}_{\text{next}} = \mathbf{x} - \eta \cdot \mathbf{g},\ \mathbf{g} \in \partial f(\mathbf{x})$.
  - ▷ **Convergence.** Assume that f is convex, minimizer $\mathbf{x}^*$ exists, Lipschitz condition ($\|\mathbf{g}\| \leq M, \forall \mathbf{g}, \mathbf{x}$) holds, initialization $\mathbf{x}^{(1)}$ satisfies $\|\mathbf{x}^{(1)} - \mathbf{x}^*\| \leq R$ for some finite R. Using subgradient method with any sequence of step sizes $\{\eta_k\}_{k=1}^{\infty}$ satisfying $\lim_{k\to\infty} \eta_k = 0$ and $\sum_{k=1}^{\infty} \eta_k = \infty$, we have as $k \to \infty$
  $$\min_{k=1,\cdots,K} f\left(\mathbf{x}^{(k)}\right) \to f(\mathbf{x}^*).$$
    - ∗ Choosing $\eta_k = \frac{\eta_0}{\sqrt{k}}$, we yield a convergence rate of $O(\frac{\log k}{\sqrt{k}})$.

Proof. How close the $(k+1)$-th iterate is to $\mathbf{x}^*$?

$$\begin{aligned}\frac{1}{2}\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\right\|^2 &= \frac{1}{2}\left\|\mathbf{x}^{(k)} - \eta_k \mathbf{g}^{(k)} - \mathbf{x}^*\right\|^2 \\ &= \frac{1}{2}\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\|^2 - \eta_k \mathbf{g}^{(k)\top}\left(\mathbf{x}^{(k)} - \mathbf{x}^*\right) + \frac{\eta_k^2}{2}\left\|\mathbf{g}^{(k)}\right\|^2 \\ &\leq \frac{1}{2}\left\|\mathbf{x}^{(k)} - \mathbf{x}^*\right\|^2 - \eta_k\left(f\left(\mathbf{x}^{(k)}\right) - f\left(\mathbf{x}^*\right)\right) + \frac{\eta_k^2}{2}\left\|\mathbf{g}^{(k)}\right\|^2.\end{aligned}$$

Rearranging and summing from $k = 1$ to K:

$$\begin{aligned}\sum_{k=1}^{K} \eta_k\left(f\left(\mathbf{x}^{(k)}\right) - f\left(\mathbf{x}^*\right)\right) &\leq \frac{1}{2}\left\|\mathbf{x}^{(1)} - \mathbf{x}^*\right\|^2 - \frac{1}{2}\left\|\mathbf{x}^{(K+1)} - \mathbf{x}^*\right\|^2 + \sum_{k=1}^{K}\frac{\eta_k^2}{2}\left\|\mathbf{g}^{(k)}\right\|^2 \\ &\leq \frac{1}{2}\left\|\mathbf{x}^{(1)} - \mathbf{x}^*\right\|^2 + \sum_{k=1}^{K}\frac{\eta_k^2}{2}\left\|\mathbf{g}^{(k)}\right\|^2 \\ &\leq \frac{1}{2}R^2 + \frac{1}{2}M^2\sum_{k=1}^{K}\eta_k^2.\end{aligned}$$

$$\min_{k=1,\cdots,K} f\left(\mathbf{x}^{(k)}\right) - f\left(\mathbf{x}^*\right) \leq \frac{\frac{1}{2}R^2 + \frac{1}{2}M^2\sum_{k=1}^{K}\eta_k^2}{\sum_{k=1}^{K}\eta_k} \to 0 \text{ as } K \to \infty.$$

Projected Gradient-Based Optimization: Constrained problems.

- Projected gradient descent: $\mathbf{x}_{\text{next}} = \Pi_{\mathcal{C}}(\mathbf{x} - \eta \cdot \mathbf{g})$.
  - ▷ Projected to the closest point in feasible set $\mathcal{C}$.

7 Boosting

Decision Stump: $h(\mathbf{x}, \boldsymbol{\theta}) = h(\mathbf{x}, \{s, k, \theta_0\}) = \text{sign}(s(x_k - \theta_0))$.

- Choose $s \in \{1, -1\}$ s.t. h is $> \frac{1}{2}$ correct.

AdaBoost: Weighted aggregation of simple models (decision stumps).

- Exponential loss: $(y, f(\mathbf{x})) \to \exp(-yf(\mathbf{x}))$.

① Initialize $w_0(t) = 1/n$ for $t = 1, \cdots, n$.

② For $m = 1, \cdots, M$ do

▷ Choose the next base learner $h(\cdot, \hat{\boldsymbol{\theta}}_m)$ as
$$\hat{\boldsymbol{\theta}}_m = \arg\min_{\boldsymbol{\theta}} \sum_{t: y_t \neq h(\mathbf{x}_t, \boldsymbol{\theta})} w_{m-1}(t).$$

▷ Set $\hat{\alpha}_m = \frac{1}{2}\log\frac{1-\hat{\epsilon}_m}{\hat{\epsilon}_m}$, where $\hat{\epsilon}_m = \sum_{t: y_t \neq h(\mathbf{x}_t, \hat{\boldsymbol{\theta}}_m)} w_{m-1}(t)$.

▷ Update the weights and normalize by Z_m:
$$w_m(t) = \frac{1}{Z_m} w_{m-1}(t) e^{-y_t h(\mathbf{x}_t, \hat{\boldsymbol{\theta}}_m)\hat{\alpha}_m};$$
$$Z_m = \sum_{t=1}^{n} w_{m-1}(t) e^{-y_t h(\mathbf{x}_t, \hat{\boldsymbol{\theta}}_m)\hat{\alpha}_m}.$$

③ Output $f_M(\mathbf{x}) = \sum_{m=1}^{M} \hat{\alpha}_m h(\mathbf{x}, \hat{\boldsymbol{\theta}}_m)$ w.r.t. classifier $\text{sign}(f_M(\mathbf{x}))$.

- Sum of weight w of wrongly classified examples is 1/2.
- **Convergence.** After M iterations, the training error satisfies
$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{y_t f_M(\mathbf{x}_t) \leq 0\} \leq \exp\left(-2\sum_{m=1}^{M}\left(\frac{1}{2} - \hat{\epsilon}_m\right)^2\right).$$
In particular, if $\hat{\epsilon}_m \leq \frac{1}{2} - \gamma$ for all m and some $\gamma > 0$, then
$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{y_t f_M(\mathbf{x}_t) \leq 0\} \leq \exp\left(-2M\gamma^2\right).$$

Proof. The 0-1 loss is bounded by exponential loss:

$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{y_t f_M(\mathbf{x}_t) \leq 0\} \leq \frac{1}{n}\sum_{t=1}^{n} e^{-y_t f_M(\mathbf{x}_t)} = \prod_{m=1}^{M} Z_m.$$

$$\begin{aligned}Z_m &= \sum_{t: y_t \neq h(\mathbf{x}_t, \boldsymbol{\theta})} e^{\hat{\alpha}_m} w_{m-1}(t) + \sum_{t: y_t = h(\mathbf{x}_t, \boldsymbol{\theta})} e^{-\hat{\alpha}_m} w_{m-1}(t) \\ &= e^{\hat{\alpha}_m}\hat{\epsilon}_m + e^{-\hat{\alpha}_m}(1 - \hat{\epsilon}_m) \\ &= \sqrt{\frac{1-\hat{\epsilon}_m}{\hat{\epsilon}_m}}\hat{\epsilon}_m + \sqrt{\frac{\hat{\epsilon}_m}{1-\hat{\epsilon}_m}}(1 - \hat{\epsilon}_m) \\ &= 2\sqrt{\hat{\epsilon}_m(1-\hat{\epsilon}_m)} = \sqrt{1 - (1 - 2\hat{\epsilon}_m)^2} \leq e^{-\frac{1}{2}(1-2\hat{\epsilon}_m)^2}.\end{aligned}$$

Combine the two equations above and prove our theorem:

$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{y_t f_M(\mathbf{x}_t) \leq 0\} \leq \prod_{m=1}^{M} Z_m \leq \prod_{m=1}^{M} e^{-\frac{1}{2}(1-2\hat{\epsilon}_m)^2} = e^{-2\sum_{m=1}^{M}(\frac{1}{2}-\hat{\epsilon}_m)^2}.$$

8 Statistical Learning Theory

Hoeffding's Inequality: Let $Z = X_1 + \cdots + X_n$, where $X_i \in [a_i, b_i]$:
$$\Pr\left[\frac{1}{n}|Z - \mathbb{E}[Z]| > \epsilon\right] \leq 2\exp\left(\frac{2n\epsilon^2}{\frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Empirical Risk Minimization:

- True risk: $R(f) = \mathbb{E}[\ell(y, f(\mathbf{x}))]$.
- Empirical risk: $R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \mathbf{x}_i)$.
- Test err $R(f)$ = train err $R_n(f)$ + generalization err $(R(f) - R_n(f))$.

PAC Learning: Given a loss function $\ell(\cdot, \cdot)$, a function class $\mathcal{F}$ is said to be *PAC-learnable* if there exists an algorithm $\mathcal{A}(\mathcal{D}_n)$ and a function $\bar{n}(\epsilon, \delta)$ such that for any distribution $P_{\mathbf{X}Y}$ used to generate $\mathcal{D}_n$ and any $\epsilon, \delta \in (0, 1)$, if $n > \bar{n}(\epsilon, \delta)$, the following holds with probability at least $1 - \delta$:
$$R(\hat{f}) \leq \min_{f \in \mathcal{F}} R(f) + \epsilon.$$

① The probability $1 - \delta$ corresponds to *probably correct.*

② The error ϵ corresponds to *approximately correct.*

③ The function $\bar{n}$ is called the *sample complexity.*

- **Thm.** For any bounded loss function $\ell(y, f(\mathbf{x})) \in [0, 1]$, any finite function class $\mathcal{F}$ is PAC-learnable with sample complexity
$$\bar{n}(\epsilon, \delta) = \frac{2}{\epsilon^2}\log\frac{2|\mathcal{F}|}{\delta}.$$

Proof. By Hoeffding's inequality,
$$\Pr[|R(f) - R_n(f)| \geq \epsilon_0] \leq 2e^{-2n\epsilon_0^2}.$$
Apply the union bound $\Pr[A_1 \cup \cdots \cup A_m] \leq \sum_{i=1}^{m} \Pr[A_i]$:
$$\Pr\left[\bigcup_{f \in \mathcal{F}} \{|R(f) - R_n(f)| \geq \epsilon_0\}\right] \leq 2|\mathcal{F}|e^{-2n\epsilon_0^2}.$$
Setting RHS as δ, we get $n = \frac{1}{2\epsilon_0^2}\log\frac{2|\mathcal{F}|}{\delta}$, or $\epsilon_0 = \sqrt{\frac{1}{2n}\log\frac{2|\mathcal{F}|}{\delta}}$.

Let $f^* = \arg\min_{f \in \mathcal{F}} R(f)$. With probability $1 - \delta$,
$$\begin{aligned}R(f\text{erm} - R(f^*) &= R(f\text{erm}) - R_n(f\text{erm}) + R_n(f\text{erm}) - R_n(f^*) + \\ &\quad R_n(f^*) - R(f^*) \\ &\leq \epsilon_0 + 0 + \epsilon_0 = 2\epsilon_0.\end{aligned}$$
Setting $\epsilon_0 = \epsilon/2$ and $\bar{n}(\epsilon, \delta) = \frac{2}{\epsilon^2}\log\frac{2|\mathcal{F}|}{\delta}$, we have proven.

▷ **Col.** With probability at least $1 - \delta$, for all $f \in \mathcal{F}$ we have:
$$|R(f) - R_n(f)| \leq \frac{1}{2n}\log\frac{2|\mathcal{F}|}{\delta}.$$

▷ **Col.** With probability at least $1 - \delta$,
$$R(f\text{erm}) - \min_{f \in \mathcal{F}} R(f) \leq \sqrt{\frac{2}{n}\log\frac{2|F|}{\delta}}.$$

Infinite Hypothesis Class:

- Growth function: Given any n unlabelled data, how many different assignments of labels can functions in $\mathcal{F}$ make?
$$S_n(\mathcal{F}) = \sup_{\mathbf{x}_1, \cdots, \mathbf{x}_n} |\{(f(\mathbf{x}_1), \cdots, f(\mathbf{x}_n)) : f \in \mathcal{F}\}|.$$
- VC-dimension: Largest k such that $S_k(\mathcal{F}) = 2^k$ (can be ∞).
  - ▷ If $\mathcal{F}$ can provide 2^k different assignments to a set of k points $\mathbf{x}_1, \cdots, \mathbf{x}_k$, we say these k points are *shattered* by $\mathcal{F}$.
  - ▷ Sauer's lemma: $S_n(\mathcal{F}) \leq \sum_{i=0}^{d_{\text{VC}}} \binom{n}{i}$.
    - ∗ $S_n(\mathcal{F}) \begin{cases} = 2^n & n \leq d_{\text{VC}}; \\ \leq (\frac{d_{\text{VC}}e}{n})^{d_{\text{VC}}} & n > d_{\text{VC}}. \end{cases}$
  - ▷ If $d_{\text{VC}}(\mathcal{F}) < \infty$, then $\mathcal{F}$ is PAC-learnable under 0-1 loss with sample complexity $\bar{n}(\epsilon, \delta) = C \cdot (d_{\text{VC}} + \log\frac{1}{\delta})/(\epsilon^2)$ for some constant C. If $d_{\text{VC}} = \infty$, then $\mathcal{F}$ is not PAC-learnable.

9 Unsupervised Learning

K-Means Clustering: Repeat the following 2 steps:

① Assign each point to the nearest cluster center:
$$\mathcal{D}_j = \{\mathbf{x} \in \mathcal{D} : j = \arg\min_{j'=1,\cdots,K} \|\mathbf{x} - \boldsymbol{\mu}_{j'}\|^2\}.$$

② Update cluster center to the average of points in that cluster:
$$\boldsymbol{\mu}_j = \frac{1}{|\mathcal{D}_j|}\sum_{\mathbf{x} \in \mathcal{D}_j} \mathbf{x}.$$

- The objective $\sum_{j=1}^{K}\sum_{\mathbf{x} \in \mathcal{D}_j} \|\mathbf{x} - \boldsymbol{\mu}_j\|^2$ is monotone non-increasing.

Maximum Likelihood Estimate:
$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \prod_{t=1}^{n} \Pr(\mathbf{x}_t; \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{t=1}^{n} \log\Pr(\mathbf{x}_t; \boldsymbol{\theta}).$$

Appendix

Matrix Properties:

PSD	$\forall \mathbf{x} \in \mathbb{R}^n\ [\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0]$	$\forall \lambda\ [\lambda \geq 0]$	⇔ convex
PD	$\forall \mathbf{x} \neq \mathbf{0}\ [\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0]$	$\forall \lambda\ [\lambda > 0]$	⇒ strictly convex
NSD	$\forall \mathbf{x} \in \mathbb{R}^n\ [\mathbf{x}^\top \mathbf{A}\mathbf{x} \leq 0]$	$\forall \lambda\ [\lambda \leq 0]$	⇔ concave
ND	$\forall \mathbf{x} \neq \mathbf{0}\ [\mathbf{x}^\top \mathbf{A}\mathbf{x} < 0]$	$\forall \lambda\ [\lambda < 0]$	⇒ strictly concave
ID	none of the above	$\lambda_1 > 0; \lambda_2 < 0$	⇒ neither nor

- $\mathbf{X}^\top \mathbf{X}$ is symmetric and PSD; $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}$ is PD.
- $\text{Eig}(\mathbf{A} + \mathbf{I}) = \text{Eig}(\mathbf{A}) + 1$. PSD + PD = PD.
- Product of eigenvalues is equal to determinant.
- Trace: ① linear $(\text{Tr}(\mathbb{E}[\mathbf{A}]) = \mathbb{E}[\text{Tr}(\mathbf{A})])$; ② $\mathbf{u}^\top \mathbf{v} = \text{Tr}(\mathbf{u}^\top \mathbf{v}) = \text{Tr}(\mathbf{v}^\top \mathbf{u})$.
- Derivative: $\nabla_{\mathbf{x}}\|\mathbf{A}\mathbf{x} + \mathbf{b}\|^2 = 2\mathbf{A}^\top(\mathbf{A}\mathbf{x} + \mathbf{b})$.