MA4270 Data Modelling and Computation

Midterm 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 accurately.

- 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: $f_{\boldsymbol\theta} = \operatorname{sign}(\boldsymbol\theta^\top \mathbf{x})$.
- Training error: $\hat{E}(\boldsymbol\theta) = \frac{1}{n}\sum_{t=1}^n \operatorname{Loss}(y_t, f_{\boldsymbol\theta}(\mathbf{x}_t))$.
  - ▷ $\operatorname{Loss}(y,\hat{y}) = \mathbf{1}\{\hat{y} \neq y\} = \begin{cases} 1 & \hat{y}\neq y \\ 0 & \text{otherwise}\end{cases}$.
  - ▷ A dataset is *linearly separable* if $\exists \boldsymbol\theta\ \left[\hat{E}(\boldsymbol\theta) = 0\right]$.

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: $\exists R \in (0,\infty)\ \forall \mathbf{x}_t \in \mathcal{D}\ [\|\mathbf{x}_t\| \le R]$.
  (2) Linearly separable: $\exists \boldsymbol\theta^*\ \exists\gamma > 0\ \left[\min_{t=1,2,\cdots,n} y_t\left(\boldsymbol\theta^*\right)^\top \mathbf{x}_t \ge \gamma\right]$.
- 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 $(\boldsymbol\theta^*)^\top\boldsymbol\theta^{(k)} = (\boldsymbol\theta^*)^\top\left(\boldsymbol\theta^{(k-1)} + y_t\mathbf{x}_t\right) \ge (\boldsymbol\theta^*)^\top\boldsymbol\theta^{(k-1)} + \gamma$. So $(\boldsymbol\theta^*)^\top\boldsymbol\theta^{(k)} \ge k\gamma$.
② $\|\boldsymbol\theta^{(k)}\|^2 = \|\boldsymbol\theta^{(k-1)}\|^2 + 2\langle\boldsymbol\theta^{(k-1)}, y_t\mathbf{x}_t\rangle + \|\mathbf{x}_t\|^2 \le \|\boldsymbol\theta^{(k-1)}\|^2 + \|\mathbf{x}_t\|^2$. So $\|\boldsymbol\theta^{(k)}\|^2 \le kR^2$.
③ By Cauchy-Schwarz inequality $\langle\mathbf{v},\mathbf{w}\rangle \le \|\mathbf{v}\|\cdot\|\mathbf{w}\|$, we have $1 \ge \frac{\langle\boldsymbol\theta^{(k)},\boldsymbol\theta^*\rangle}{\|\boldsymbol\theta^{(k)}\|\cdot\|\boldsymbol\theta^*\|} \ge \frac{k\gamma}{\sqrt{kR^2}\|\boldsymbol\theta^*\|}$, hence $k \le \frac{R^2\|\boldsymbol\theta^*\|^2}{\gamma^2}$.

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

- SVM with offset: $\min_{\boldsymbol\theta,\theta_0} \frac{1}{2}\|\boldsymbol\theta\|^2$ s.t. $\forall t\ \left[y_t\left(\boldsymbol\theta^\top\mathbf{x}_t + \theta_0\right) \ge 1\right]$.
  - ▷ 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. $\forall t\ \left[y_t\left(\boldsymbol\theta^\top\mathbf{x}_t + \theta_0\right) \ge 1 - \zeta_t\right]$.
  - ▷ $\boldsymbol\zeta = (\zeta_1, \zeta_2, \cdots, \zeta_n) \ge \mathbf{0}$ 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: $\Pr(y = c \mid \mathbf{x}) = \frac{\exp\left(\boldsymbol\theta_c^\top\mathbf{x}+\theta_{0,c}\right)}{\sum_{c'=1}^M \exp\left(\boldsymbol\theta_{c'}^\top\mathbf{x}+\theta_{0,c'}\right)}$.
  - ▷ 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: $\operatorname{Cov}[\hat{\boldsymbol\Theta}] = \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$.
- 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 $\text{bias} = \|\boldsymbol\mu\|^2 - 2\langle\boldsymbol\mu,\boldsymbol\Theta^*\rangle + \|\boldsymbol\Theta^*\|^2$.
② $\text{variance} = \mathbb{E}[\|\hat{\boldsymbol\Theta}\|^2 - 2\langle\hat{\boldsymbol\Theta},\boldsymbol\mu\rangle + \|\boldsymbol\mu\|^2] = \mathbb{E}[\|\hat{\boldsymbol\Theta}\|^2] - 2\langle\mathbb{E}[\hat{\boldsymbol\theta}],\boldsymbol\mu\rangle + \|\boldsymbol\mu\|^2 = \mathbb{E}[\|\hat{\boldsymbol\Theta}\|^2] - \|\boldsymbol\mu\|^2$.
③ $\text{bias} + \text{variance} = \mathbb{E}[\|\hat{\boldsymbol\Theta}\|^2] - 2\langle\boldsymbol\mu,\boldsymbol\Theta^*\rangle + \|\boldsymbol\Theta^*\|^2 = \text{LHS}$.

Appendix

Matrix Properties:

PSD	$\forall\mathbf{x}\in\mathbb{R}^n\ [\mathbf{x}^\top\mathbf{A}\mathbf{x} \ge 0]$	$\forall\lambda\ [\lambda \ge 0]$	⇔ convex
PD	$\forall\mathbf{x}\neq\mathbf{0}\ [\mathbf{x}^\top\mathbf{A}\mathbf{x} > 0]$	$\forall\lambda\ [\lambda > 0]$	⇒ strictly convex
NSD	$\forall\mathbf{x}\in\mathbb{R}^n\ [\mathbf{x}^\top\mathbf{A}\mathbf{x} \le 0]$	$\forall\lambda\ [\lambda \le 0]$	⇔ 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.
- $\operatorname{Eig}(\mathbf{A}+\mathbf{I}) = \operatorname{Eig}(\mathbf{A}) + 1$. PSD + PD = PD.
- Trace: ① linear ($\operatorname{Tr}(\mathbb{E}[\mathbf{A}]) = \mathbb{E}[\operatorname{Tr}(\mathbf{A})]$); ② $\mathbf{u}^\top\mathbf{v} = \operatorname{Tr}(\mathbf{u}^\top\mathbf{v}) = \operatorname{Tr}(\mathbf{v}^\top\mathbf{u})$.
- Derivative: $\nabla_{\mathbf{x}}\|\mathbf{A}\mathbf{x}+\mathbf{b}\|^2 = 2\mathbf{A}^\top(\mathbf{A}\mathbf{x}+\mathbf{b})$.