HW 3

William lee WL72

October 2025

1. Support Vector Machines

Soft-margin SVM (hinge loss) primal.

$$\min_{w,b,\{\xi_i\}} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{N} \xi_i \quad \text{s.t.} \quad y^{(i)}\big(w^\top x^{(i)} + b\big) \geq 1 - \xi_i, \quad \xi_i \geq 0.$$

Data in R^2:

$$x^{(1)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},\ y^{(1)} = +1; \qquad x^{(2)} = \begin{bmatrix} 2 \\ 0 \end{bmatrix},\ y^{(2)} = +1; \qquad x^{(3)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix},\ y^{(3)} = -1.$$

Definitions: $f(x) = w^\top x + b$; $\gamma_i := y^{(i)} f\big(x^{(i)}\big)$; $\xi_i := \max\{0,\, 1 - \gamma_i\}$.

(a) Soft margin with hinge loss.

(i) Given $w = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $b = 0$, $C = 1$: compute $f(x^{(i)})$, γ_i, ξ_i, objective

$$\begin{aligned} f\big(x^{(1)}\big) &= 1\cdot 0 + 0\cdot 0 + 0 = 0, \\ f\big(x^{(2)}\big) &= 1\cdot 2 + 0\cdot 0 + 0 = 2, \\ f\big(x^{(3)}\big) &= 1\cdot 1 + 0\cdot 1 + 0 = 1. \end{aligned} \quad \Rightarrow \quad \begin{aligned} &\gamma_1 = (+1)\cdot 0 = 0, \qquad \gamma_2 = (+1)\cdot 2 = 2, \qquad \gamma_3 = (-1)\cdot 1 = -1; \\ &\xi_1 = \max(0, 1-0) = 1, \quad \xi_2 = \max(0, 1-2) = 0, \quad \xi_3 = \max(0, 1-(-1)) = 2. \end{aligned}$$

$$\|w\|^2 = 1^2 + 0^2 = 1 \quad \Rightarrow \quad \text{objective} = \frac{1}{2}\|w\|^2 + C\sum_i \xi_i = \frac{1}{2}\cdot 1 + 1\cdot(1+0+2) = \boxed{3.5}.$$

(ii) Given $w = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$, $b = 0$, $C = 1$: repeat (i) and compare

$$\begin{aligned} &f\big(x^{(1)}\big) = 0, \qquad f\big(x^{(2)}\big) = 1\cdot 2 + (-1)\cdot 0 = 2, \qquad f\big(x^{(3)}\big) = 1\cdot 1 + (-1)\cdot 1 = 0; \\ &\gamma_1 = (+1)\cdot 0 = 0,\ \ \gamma_2 = (+1)\cdot 2 = 2,\ \ \gamma_3 = (-1)\cdot 0 = 0; \\ &\xi_1 = \max(0, 1-0) = 1,\ \ \xi_2 = \max(0, 1-2) = 0,\ \ \xi_3 = \max(0, 1-0) = 1. \end{aligned}$$

$$\|w\|^2 = 1^2 + (-1)^2 = 2 \quad \Rightarrow \quad \text{objective} = \frac{1}{2}\cdot 2 + (1+0+1) = \boxed{3.0}.$$

Comparison: $3.0 < 3.5$, so (ii) has the smaller objective.

(iii) Change C: evaluate objectives for (i) and (ii) with $C = 0.5$ and $C = 2$; discuss trade-off

Case (i): $\|w\|^2 = 1$, $\sum \xi_i = 3$.

$$C = 0.5: \ \frac{1}{2} \cdot 1 + 0.5 \cdot 3 = 2.0; \qquad C = 2: \ \frac{1}{2} \cdot 1 + 2 \cdot 3 = 6.5.$$

Case (ii): $\|w\|^2 = 2$, $\sum \xi_i = 2$.

$$C = 0.5: \ \frac{1}{2} \cdot 2 + 0.5 \cdot 2 = 2.0; \qquad C = 2: \ \frac{1}{2} \cdot 2 + 2 \cdot 2 = 5.0.$$

Trade-off: Increasing C emphasizes minimizing violations ($\sum \xi_i$), often accepting a larger $\|w\|$ (smaller margin). Decreasing C tolerates more violations to keep $\|w\|$ small (larger margin).

(b) Importance weighted soft-margin SVMs.

We are given weights $p^{(i)} \in [0, 1]$.

(i) Primal with per-example importance weights

$$\boxed{\min_{w,b,\{\xi_i\}} \ \frac{1}{2}\|w\|_2^2 + C \sum_{i=1}^{N} p^{(i)} \xi_i \quad \text{s.t.} \quad y^{(i)}(w^\top x^{(i)} + b) \geq 1 - \xi_i, \ \ \xi_i \geq 0.}$$

(ii) Dual derivation; effect of $p^{(i)}$

Lagrangian with multipliers $\alpha_i \geq 0$ (margin constraints) and $\mu_i \geq 0$ ($\xi_i \geq 0$):

$$\mathscr{L} = \tfrac{1}{2}\|w\|^2 + C \sum_i p^{(i)} \xi_i - \sum_i \alpha_i \big(y^{(i)}(w^\top x^{(i)} + b) - 1 + \xi_i \big) - \sum_i \mu_i \xi_i.$$

Stationarity:

$$\partial_w \mathscr{L} = 0 \Rightarrow w = \sum_i \alpha_i y^{(i)} x^{(i)}, \qquad \partial_b \mathscr{L} = 0 \Rightarrow \sum_i \alpha_i y^{(i)} = 0,$$

$$\partial_{\xi_i} \mathscr{L} = 0 \Rightarrow C p^{(i)} - \alpha_i - \mu_i = 0.$$

Primal feas.: $y^{(i)}(w^\top x^{(i)} + b) \geq 1 - \xi_i$, $\xi_i \geq 0$. Dual feas.: $\alpha_i \geq 0$, $\mu_i \geq 0$. Compl. slackness: $\alpha_i \big(y^{(i)}(w^\top x^{(i)} + b) - 1 + \xi_i \big) = 0$, $\mu_i \xi_i = 0$.

Eliminate w, b, μ, ξ to get the dual:

$$\boxed{\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j \, y^{(i)} y^{(j)} \, x^{(i)\top} x^{(j)} \\ \text{s.t.} \quad & \sum_{i=1}^{N} \alpha_i y^{(i)} = 0, \qquad 0 \leq \alpha_i \leq C \, p^{(i)} \ \ \forall i. \end{aligned}}$$

Thus $p^{(i)}$ tightens the upper bound from C to $C \, p^{(i)}$.

(iii) Bounds for $p^{(1)} = 1$, $p^{(2)} = \frac{1}{2}$, $p^{(3)} = 0$ with $C = 2$

$$0 \le \alpha_1 \le 2, \qquad 0 \le \alpha_2 \le 1, \qquad 0 \le \alpha_3 \le 0 \;\Rightarrow\; \alpha_3 = 0.$$

Dual equality (using $y_1 = y_2 = +1$, $y_3 = -1$):

$$\alpha_1 + \alpha_2 - \alpha_3 = 0 \overset{\alpha_3=0}{\Longrightarrow} \alpha_1 + \alpha_2 = 0.$$

With $\alpha_1, \alpha_2 \ge 0$, the only feasible solution is

$$\boxed{(\alpha_1, \alpha_2, \alpha_3) = (0, 0, 0).}$$

(iv) Dual for the L_2-slack SVM

Primal:

$$\min_{w,b,\{\xi_i\}} \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_i \xi_i^2 \quad \text{s.t.} \quad y^{(i)}(w^\top x^{(i)} + b) \ge 1 - \xi_i, \;\; \xi_i \ge 0.$$

Lagrangian:

$$\mathscr{L} = \tfrac{1}{2}\|w\|^2 + \tfrac{C}{2}\sum_i \xi_i^2 - \sum_i \alpha_i \big(y^{(i)}(w^\top x^{(i)} + b) - 1 + \xi_i\big) - \sum_i \mu_i \xi_i.$$

Stationarity:

$$w = \sum_i \alpha_i y^{(i)} x^{(i)}, \qquad \sum_i \alpha_i y^{(i)} = 0, \qquad C\xi_i - \alpha_i - \mu_i = 0.$$

Eliminating ξ, μ yields the dual (no upper box; extra quadratic penalty):

$$\boxed{\max_{\alpha \ge 0,\; \alpha^\top y = 0} \quad \mathbf{1}^\top \alpha - \tfrac{1}{2}\,\alpha^\top \Big(YKY + \tfrac{1}{C} I\Big)\alpha,}$$

with $K_{ij} = x^{(i)\top} x^{(j)}$ and $Y = \mathrm{diag}(y^{(1)}, \ldots, y^{(N)})$.

Problem 2: Implementing Support Vector Machine

Part (a): Projections for hard-/soft-margin dual domains

Sets and projection. Let the feasible sets be

$$\mathscr{C}_{\text{hard}} = [0, \infty)^N, \qquad \mathscr{C}_{\text{soft}} = [0, C]^N,$$

and define the Euclidean projection onto a closed convex set $\mathscr{C}$ by

$$\Pi_{\mathscr{C}}(a) \;=\; \arg\min_{u \in \mathscr{C}} \; \|u - a\|_2^2.$$

Claims. For any $a \in R^N$,

$$\boxed{\Pi_{[0,\infty)^N}(a)_i = \max\{a_i, 0\}}, \qquad \boxed{\Pi_{[0,C]^N}(a)_i = \min\big\{\max\{a_i, 0\}, C\big\}}.$$

Proof sketch (componentwise).

1. Both $\mathscr{C}_{\text{hard}}$ and $\mathscr{C}_{\text{soft}}$ are Cartesian products of 1D convex sets: $[0,\infty)$ and $[0,C]$. The squared distance separates:

$$\|u-a\|_2^2 \;=\; \sum_{i=1}^{N}(u_i-a_i)^2.$$

2. Hence the projection decouples into N independent 1D problems:

$$\min_{u_i\in[0,\infty)}(u_i-a_i)^2 \quad \text{or} \quad \min_{u_i\in[0,C]}(u_i-a_i)^2.$$

3. For $[0,\infty)$: if $a_i \geq 0$ the minimizer is $u_i = a_i$; if $a_i < 0$ the closest feasible point is $u_i = 0$. This yields $u_i = \max\{a_i, 0\}$.

4. For $[0,C]$: clip a_i into the interval endpoints. If $a_i < 0$ use 0; if $0 \leq a_i \leq C$ use a_i; if $a_i > C$ use C. Thus $u_i = \min\{\max(a_i, 0), C\}$.

5. Uniqueness follows because the sets are closed and convex and the objective is strictly convex componentwise.

Part (b): `svm_solver()` via Projected Gradient Descent (PGD)

Dual (minimization form). Let $X \in R^{N\times d}$ have rows $x_i^\top$, labels $y \in \{-1,+1\}^N$, Gram matrix K with $K_{ij} = x_i^\top x_j$ (or a kernel), and $Y = \mathrm{diag}(y)$. Define $Q := YKY$ (which is PSD). We minimize

$$\boxed{f(\alpha) \;=\; \tfrac{1}{2}\,\alpha^\top Q\,\alpha \;-\; \mathbf{1}^\top\alpha} \quad \text{over} \quad \alpha \in \begin{cases} [0,C]^N & \text{(soft margin)},\\ [0,\infty)^N & \text{(hard margin)}, \end{cases}$$

with gradient

$$\boxed{\nabla f(\alpha) \;=\; Q\,\alpha \;-\; \mathbf{1}}.$$

PGD update. With step size $\eta > 0$,

$$\boxed{\alpha_{t+1} \;=\; \Pi_{\mathscr{C}}(\alpha_t - \eta\,(Q\alpha_t - \mathbf{1}))\,,}$$

where $\mathscr{C} = [0,C]^N$ for soft margin (use `clamp(0,C)`) and $\mathscr{C} = [0,\infty)^N$ for hard margin (use `clamp(0, +∞)`).

Algorithmic steps.

- Inputs: $X \in R^{N\times d}$, $y \in \{-1,+1\}^N$, C (finite for soft, $+\infty$ for hard), step size η, iterations T, tolerance tol.
- Precompute: $K = XX^\top$ (or kernel matrix), $Q = YKY$.
- Initialize: $\alpha^{(0)} = \mathbf{0}$.
- Loop for $t = 0,\dots,T-1$:

$$g_t = Q\alpha^{(t)} - \mathbf{1}, \qquad z_t = \alpha^{(t)} - \eta\, g_t, \qquad \alpha^{(t+1)} = \Pi_{\mathscr{C}}(z_t).$$

  Stop if $\|\alpha^{(t+1)} - \alpha^{(t)}\|_2 \le \texttt{tol}$.

- Linear kernel outputs:

$$\alpha_\star = \alpha^{(t)}, \qquad w = \sum_{i=1}^{N} \alpha_{\star,i}\, y_i\, x_i \quad \text{(vectorized: } (\alpha_\star \odot y)^\top X\text{)}.$$

Torch notes. Use torch.matmul for products and torch.clamp for projection. You may enclose the projection/update inside torch.no_grad() since the analytic gradient is $Q\alpha - \mathbf{1}$. If experimenting with backward(), remember to zero grads.

Complexity. Each PGD iteration costs $O(N^2)$ for the matrix{vector product $Q\alpha$ (if Q is explicit). Memory is $O(N^2)$ if you store Q.

Part (c): svm_predictor() and computation of the bias b

Linear kernel: computing w.

$$w = \sum_{i=1}^{N} \alpha_i\, y_i\, x_i.$$

Compute b from a specific support vector.

- Select the support index $i^\star$ as the *minimum positive* α_i (soft: $0 < \alpha_i < C$; hard: $\alpha_i > 0$).
- By KKT complementary slackness for such an $i^\star$,

$$y_{i^\star}\left(w^\top x_{i^\star} + b\right) = 1 \quad \Rightarrow \quad \boxed{b = y_{i^\star} - w^\top x_{i^\star}}.$$

- Numerical note: if there is no exact $0 < \alpha_i < C$ due to tolerances, pick the smallest α_i in $(\varepsilon, C-\varepsilon)$.

Prediction (return raw scores). For any $x \in R^d$,

$$f(x) = w^\top x + b,$$

and for a batch $Z \in R^{m\times d}$,

$$\texttt{scores} = Zw + b\mathbf{1}_m.$$

Return the raw value $f(x)$ (do not map to ± 1).

Kernelized decision function (optional). With kernel $K(\cdot,\cdot)$,

$$f(x) = \sum_{j=1}^{N} \alpha_j y_j K(x_j, x) + b,$$

and the same b formula using the selected support vector $i^\star$:

$$b = y_{i^\star} - \sum_{j=1}^{N} \alpha_j y_j K(x_j, x_{i^\star}).$$

3. Linear Regression and ERM

(a) Robustness of Linear Regression (1D), with w_0 fixed to 1

Setup. Model $y \approx w_1 x + w_0$ with the constraint $w_0 = 1$.

L2 loss: $L_2(w_1) = \sum_i \big(y_i - (w_1 x_i + 1)\big)^2$, L1 loss: $L_1(w_1) = \sum_i \big|y_i - (w_1 x_i + 1)\big|$.

(i) L2 (no outlier): data $\{(1,2),(2,3),(3,6),(4,7),(5,10)\}$. Closed form with fixed intercept:

$$\boxed{w_1 = \frac{\sum_i x_i (y_i - 1)}{\sum_i x_i^2}}.$$

Using $\sum_i x_i^2 = 55$ and $\sum_i x_i (y_i - 1) = 178$,

$$w_1 = \frac{178}{110} = \boxed{1.6181818182}.$$

Check (normal equation): $\frac{d}{dw_1} L_2(w_1) = 2\sum_i x_i (w_1 x_i + 1 - y_i) = 0$ at the optimum.

(ii) L2 (with outlier): data $\{(1,2),(2,3),(3,6),(4,7),(5,10),(6,180)\}$.

$$\sum_i x_i^2 = 91, \qquad \sum_i x_i (y_i - 1) = 2326 \quad \Longrightarrow \quad w_1 = \frac{2326}{182} = \boxed{12.7802197802}.$$

Interpretation: a single extreme outlier drives the L2 slope to a very large value (non-robust).

(iii) L1 (with the same outlier), still $w_0 = 1$. Let $r_i(w_1) = y_i - (w_1 x_i + 1)$. Optimality (subgradient balance):

$$0 \in \partial L_1(w_1) \;=\; -\sum_i \operatorname{sign}(r_i(w_1))\, x_i, \quad \texttt{with} \;\; \operatorname{sign}(0) \in [-1,1].$$

Evaluate at breakpoints $w_1 = (y_i - 1)/x_i$ and locate where the subgradient crosses 0. The minimizer is

$$\boxed{w_1^{=1.8,} \qquad L_1(w_1^{)=172.2}}.$$

Interpretation: L1 is robust; the slope follows the majority trend rather than the outlier.

(b) Lasso Regression with $X^\top X = I$ (and $w_0 = 0$)

Objective.

$$\min_{w \in R^d} \;\; \|y - Xw\|_2^2 + \lambda \|w\|_1, \qquad X^\top X = I.$$

Let X_i be column i of X and $a_i := X_i^\top y$.

Derivation.

$$\|y - Xw\|_2^2 = y^\top y - 2w^\top X^\top y + w^\top X^\top X w = \texttt{const} + \sum_i \big(w_i^2 - 2a_i w_i\big).$$

Thus the problem is separable:

$$\min_{w_i} \; \phi_i(w_i) := w_i^2 - 2a_i w_i + \lambda |w_i|.$$

By subgradient conditions,

$$w_i = \begin{cases} a_i - \frac{\lambda}{2}, & a_i > \frac{\lambda}{2}, \\ 0, & |a_i| \le \frac{\lambda}{2}, \\ a_i + \frac{\lambda}{2}, & a_i < -\frac{\lambda}{2}, \end{cases} \quad \iff \quad \boxed{w_i^{=\operatorname{sign}(a_i)\,\max\{|a_i| - \frac{\lambda}{2}, 0\}}}.$$

Answers to subparts. (i) Under $X^\top X = I$, w_i^* depends only on $a_i = X_i^\top y$ and λ. (ii) If $w_i^{>0}$ then $w_i^{=a_i - \lambda/2}$ with $a_i > \lambda/2$. (iii) If $w_i^{<0}$ then $w_i^{=a_i + \lambda/2}$ with $a_i < -\lambda/2$. (iv) $w_i^{=0}$ iff $|a_i| \le \lambda/2$ (weakly correlated features are pruned).

(c) Ridge Regression (centered y and x)

Objective (centered data). Assume $\sum_i y_i = 0$ and $\sum_i x^{(i)} = \mathbf{0}$. Solve

$$\min_{w,w_0} \ \frac{1}{N}\sum_{i=1}^{N}\left(y_i - w^\top x_i - w_0\right)^2 + \lambda\|w\|_2^2.$$

Centering implies $w_0^{=0}$. With $X \in R^{N\times d}$ (rows $x_i^\top$) and $y \in R^N$:

$$\boxed{w^{=\left(\frac{1}{N}X^\top X+\lambda I\right)^{-1}\left(\frac{1}{N}X^\top y\right)}, \qquad w_0^{=0.}}$$

Warm-up: $d = 1$. Let $s_{xx} = (1/N)\sum_i x_i^2$ and $s_{xy} = (1/N)\sum_i x_i y_i$. Then

$$(1/N)\sum_i (y_i - wx_i)^2 + \lambda w^2 = (s_{xx}+\lambda)w^2 - 2s_{xy}w + \texttt{const},$$

so the FOC gives

$$\boxed{w^{=\frac{s_{xy}}{s_{xx}+\lambda}}, \qquad w_0^{=0.}}$$

(Equivalently $w^{=\frac{\sum_i x_i y_i}{\sum_i x_i^2 + N\lambda}}$ using unnormalized sums.)

General case (d ¿ 1): Ridge Regression with an intercept

Problem. Given $X \in R^{N\times d}$ (rows are $x_i^\top$), $y \in R^N$, ridge parameter $\lambda \geq 0$, and $\mathbf{1} \in R^N$ (all-ones), solve

$$\min_{w\in R^d,\, w_0 \in R} \ \frac{1}{N}\left\| y - Xw - w_0\mathbf{1}\right\|_2^2 \ + \ \lambda\|w\|_2^2,$$

where only w is penalized (the intercept w_0 is unregularized).

Useful notation. Let

$$\bar{x} \ := \ \frac{1}{N}X^\top\mathbf{1} \in R^d, \qquad \bar{y} \ := \ \frac{1}{N}\mathbf{1}^\top y \in R, \qquad H \ := \ I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top.$$

Define centered data

$$X_c \ := \ HX \ = \ X - \mathbf{1}\bar{x}^\top, \qquad y_c \ := \ Hy \ = \ y - \bar{y}\mathbf{1}.$$

Eliminate the intercept. The derivative w.r.t. w_0 gives

$$-\frac{2}{N}\mathbf{1}^\top\left(y - Xw - w_0\mathbf{1}\right) = 0 \implies w_0^{= \bar{y} - \bar{x}^\top w.}$$

Substitute back into the residual:

$$r \ = \ y - Xw - w_0^{\mathbf{1} = (y-\bar{y}\mathbf{1}) - (X-\mathbf{1}\bar{x}^\top)w = y_c - X_c w.}$$

Hence the problem reduces to

$$\min_{w\in R^d} \ \frac{1}{N}\|y_c - X_c w\|_2^2 + \lambda\|w\|_2^2.$$

Normal equations and closed form. First-order optimality w.r.t. w yields

$$\frac{1}{N}X_c^\top(X_c w - y_c) + \lambda w = 0 \implies \left(\frac{1}{N}X_c^\top X_c + \lambda I\right)w = \frac{1}{N}X_c^\top y_c.$$

Therefore

$$\boxed{w = \left(\tfrac{1}{N}X_c^\top X_c + \lambda I\right)^{-1}\left(\tfrac{1}{N}X_c^\top y_c\right), \qquad w_0 = \bar{y} - \bar{x}^\top w.}$$

Equivalent block normal equations (no centering). Solving the $(d+1)\times(d+1)$ system directly gives the same solution:

$$\begin{bmatrix} \frac{1}{N}X^\top X + \lambda I & \frac{1}{N}X^\top \mathbf{1} \\ \frac{1}{N}\mathbf{1}^\top X & 1 \end{bmatrix} \begin{bmatrix} w \\ w_0 \end{bmatrix} = \begin{bmatrix} \frac{1}{N}X^\top y \\ \bar{y} \end{bmatrix}.$$

Properties and implementation notes.

- H is symmetric idempotent ($H = H^\top = H^2$) and $H\mathbf{1} = 0$, so centering removes the need to fit an intercept inside the penalized system.
- Prefer solving the linear system $Aw = b$ with $A = \frac{1}{N}X_c^\top X_c + \lambda I$ and $b = \frac{1}{N}X_c^\top y_c$, rather than forming A^{-1} explicitly.
- Sanity checks: if data are already centered ($\bar{x} = 0$, $\bar{y} = 0$), then $w_0 = 0$ and

  $$w = \left(\tfrac{1}{N}X^\top X + \lambda I\right)^{-1}\left(\tfrac{1}{N}X^\top y\right).$$

  As $\lambda \to 0$, this reduces to ordinary least squares on the centered data.

4. Implementing Linear Regression

(a) Data Preparation & OLS Baseline

Split (70/15/15). Given features X and targets y, perform a stratified (if needed) or random split with fixed seed: first split train vs. temp (70% vs. 30%), then split temp equally into validation and test (15% / 15%).

Preprocessing pipeline.

1. Identify dtypes: partition columns into numerical vs. categorical.
2. Impute (train stats only): numerical $\to$ train median; categorical $\to$ train mode. Apply these train statistics to train/val/test.
3. One-hot encode categorical with train columns as template; align val/test by reindexing missing dummies to 0.

4. Z-score on numerical using train mean μ and std $\sigma > 0$:

$$z = \frac{x - \mu_{\mathtt{train}}}{\sigma_{\mathtt{train}}} \quad \text{(with } \sigma = 1 \text{ fallback for zero variance).}$$

5. Concatenate numerical (z-scored) + one-hot categorical, cast to float64, and prepend a bias column of ones for linear models.

OLS (normal equation). With bias included in $X_b \in R^{N\times(d+1)}$, the closed form is

$$w_{\mathtt{OLS}} = \operatorname{pinv}(X_b)\, y, \qquad \hat{y} = X_b\, w_{\mathtt{OLS}}.$$

Report MSE$= \frac{1}{N}\|\hat{y} - y\|_2^2$ and RMSE$= \sqrt{\text{MSE}}$.

(b) Ridge Regression (L2 / MAP) with Unpenalized Bias

Closed form with unpenalized w_0. Let $X_b = [\mathbf{1}, X]$ (bias first column). The ridge objective

$$\min_w \; \frac{1}{N}\|X_b w - y\|_2^2 + \lambda \|w_{1:}\|_2^2$$

leads to the linear system

$$\left(X_b^\top X_b + \lambda I_0\right) w = X_b^\top y, \quad \text{where} \quad I_0 = \operatorname{diag}(0, 1, \ldots, 1).$$

Thus

$$w_{\mathtt{ridge}}(\lambda) = \left(X_b^\top X_b + \lambda I_0\right)^{-1} X_b^\top y, \qquad \hat{y} = X_b\, w_{\mathtt{ridge}}.$$

Tune λ on validation MSE.

(c) Lasso via ISTA (proximal gradient)

Objective (bias unpenalized).

$$\min_w \; J(w) = \frac{1}{N}\|X_b w - y\|_2^2 + \lambda \|w_{1:}\|_1.$$

The smooth part $f(w) = \frac{1}{N}\|X_b w - y\|_2^2$ has gradient

$$\nabla f(w) = \frac{2}{N} X_b^\top (X_b w - y).$$

Let $\alpha \in (0, 1/L)$ with L the Lipschitz constant of ∇f (we estimate L via power iteration). ISTA update:

$$\tilde{w}^{(k+1)} = w^{(k)} - \alpha \nabla f\big(w^{(k)}\big), \qquad w_0^{(k+1)} = \tilde{w}_0^{(k+1)}, \qquad w_j^{(k+1)} = \operatorname{soft}\Big(\tilde{w}_j^{(k+1)},\; \alpha\lambda\Big) \quad (j \geq 1),$$

where $\operatorname{soft}(z, \tau) = \operatorname{sign}(z)\max(|z| - \tau, 0)$. Early stopping when $\|w^{(k+1)} - w^{(k)}\|_\infty <$ tol.

(d) Log Transform & Duan's Smearing

Motivation. SalePrice is right-skewed (heavy tail). Transform $y \mapsto y^{(\log)} = \log(1+y)$ to reduce skewness, stabilize residual variance, and improve linear-model assumptions.

Back-transform with smearing. Fit a linear model on $y^{(\log)}$ to obtain predictions $z = \hat{y}^{(\log)} = X_b w$. Let residuals on train be $r = y^{(\log)} - z$. Duan's estimator uses

$$s = \frac{1}{N}\sum_{i=1}^{N} e^{r_i} \approx E[e^r].$$

For $\log 1p$, the unbiased back-transform is

$$\widehat{y}_{\texttt{back}} = s e^z - 1.$$

Compute test predictions by evaluating z on test, multiplying by s, and subtracting 1; then report MSE/RMSE in original dollar scale.