STA 35C Statistical Data Science III

Practice final A solutions

Instructor: Dogyoon Song

Problem 1. Basic concepts

(a) **A, B and D.** Each of five fits holds out $40/5 = 8$ observations and trains on 32. Each observation appears in one validation set and four training sets.

(b) **B and D (H_2, H_4).** The ascending order is H_2, H_4, H_1, H_3. The first two p-values satisfy $0.009 \le 0.05/4$ and $0.016 \le 0.05/3$. Holm stops at the third because $0.031 > 0.05/2$.

(c) **A and D.** Single linkage takes the minimum cross-cluster distance, 2. Complete linkage takes the maximum, 7.

(d) **False.** At this x, zero bias leaves $\mathrm{Var}(Y \mid X = x) + \mathrm{Var}_{\mathrm{train}}(\widehat{f}(x))$ in the expected test error, and the second term can be positive.

(e) **True.** The lasso penalty can shrink a fitted slope to exactly zero even when the corresponding population slope is nonzero.

(f) **False.** At a fixed RSS, increasing the penalty multiplier makes a curve with greater integrated squared curvature more costly.

(g) The slope is $12/8 = 1.5$ and the intercept is $10 - 1.5 \times 4 = 4$. Thus $\boldsymbol{\widehat{y} = 4 + 1.5x}$ and $\boldsymbol{\widehat{y}(6) = 13}$.

(h) Increasing x by 1 multiplies the fitted odds by $e^{\log 2} = 2$. The new odds are $3 \times 2 = \mathbf{6}$ to 1.

(i) The score is the projection $(2, 1) \cdot (3/5, 4/5) = 6/5 + 4/5 = \mathbf{2}$.

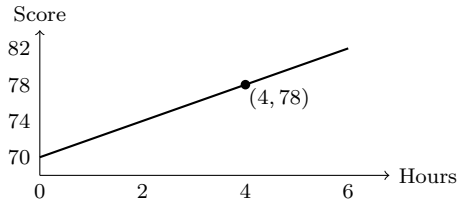
(j) Holding the number of completed quizzes fixed, one additional study hour increases the fitted exam score by 3 points.

(k) Among directions of unit length, PCA chooses one that maximizes the **sample variance** of the centered observations projected onto it.

(l) Different starting centers can assign observations to different groups at the first step, and the subsequent updates can converge to different local solutions.

Problem 2. Regression and classification

(a) **(i)** Section A's slope is **2** score points per study hour and Section B's is **1**. **(ii)** In Section A the fitted line joins $(0, 70)$ and $(6, 82)$, and the prediction at 4 hours is $70 + 2 \times 4 = \mathbf{78}$ points.



(b) **(i)** Write D for a defective part and A for an alarm. Then $P(A) = 0.8 \times 0.2 + 0.1 \times 0.8 = 0.24$. Bayes' rule gives $P(D \mid A) = 0.16/0.24 = \mathbf{2/3}$, so classify the part as **defective**. **(ii)** With prevalence p, the cutoff requires

$$\frac{0.8p}{0.8p + 0.1(1-p)} \geq 0.60 \quad \Longleftrightarrow \quad 0.38p \geq 0.06.$$

The smallest prevalence is $\mathbf{3/19 \approx 0.158}$.

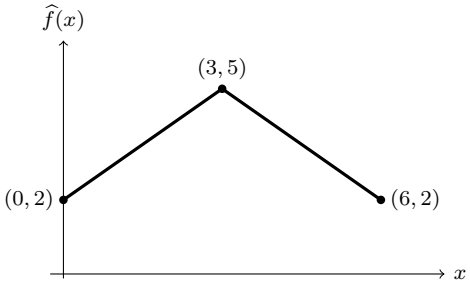
Problem 3. Model selection and the bootstrap

(a) **(i)** The original median error is **3**. The resample medians are **2, 3, 4**. **(ii)** They average to 3, so $s = \sqrt{(1+0+1)/2} = \mathbf{1}$. The bootstrap SE is $\boldsymbol{s}$ because the spread of the resample medians estimates the sampling spread of one sample median. Dividing by $\sqrt{3}$ would estimate the standard error of the average of three independent replicate medians, conditional on the original data. Three replicates give only a rough numerical SE estimate.

(b) **(i)** Forward selection first chooses X_1 (RSS 38), then compares pairs 12, 13 and 14 (RSS 26, 35 and 30) and chooses X_1, X_2. Best subset compares all six pairs and chooses X_2, X_3 (RSS 10). **(ii)** Resampling test cases leaves the fitted rule fixed. Draw training cases with replacement at the original training-sample size and rerun selection and fitting on each resample. The spread of the predictions at the same input estimates training variability.

Problem 4. Regression splines and smoothing splines

(a) **(i)** The fitted curve is $2 + x$ for $0 \leq x \leq 3$ and $8 - x$ for $3 \leq x \leq 6$.
(ii) The hinge fit has slopes $+1$ and -1 and is continuous at 3. The indicator fit has slope $+1$ on both sides. Its left and right limits are 5 and 3, so it **jumps downward by 2**.



(b) **(i)** At the repeated input, $(2-t)^2 + (4-t)^2 = 2(t-3)^2 + 2$, minimized at $t = 3$. The fitted values approach **3, 7, 6** and the smallest RSS is **2**. A smooth curve through these three means has finite roughness J_0. Optimality gives $2 \leq \text{RSS} \leq 2 + \lambda J_0$, so the residual sum above its minimum tends to zero. **(ii)** The original RSS is $2 + 2[f(1) - 3]^2 + [f(2) - 7]^2 + [f(3) - 6]^2$. Giving the three means weights **2, 1, 1** changes the objective only by the additive constant 2, hence preserves the minimizers at every λ.

Problem 5. Principal components and clustering

(a) The variances along u, v, w are $2^2 = 4$, $1^2 = 1$, and $3^2 = 9$. Thus w is the PC1 direction (up to sign). The first two PCs together explain $(9 + 4)/(4 + 1 + 9) = \mathbf{13/14}$ of the total variance.

(b) **(i)** The current centroids are $(0 + 4 + 8)/3 = 4$ and $(10 + 11)/2 = 10.5$. The boundary between them is their midpoint, 7.25. The new clusters are $\{0, 4\}$ and $\{8, 10, 11\}$. **(ii)** For fixed memberships, each group mean uniquely minimizes its sum of squared distances. The group containing 8 now has mean $(8 + 10 + 11)/3 = 29/3$, which differs from the center 10.5 used for that group in (i). Thus the original centroids no longer minimize the objective for these memberships.

(c) At $h = 1.5$, only A and B have merged, giving $\{A, B\}, \{C\}, \{D\}$. C and D merge at height 2, and the two pairs join at height 5. Thus every cut strictly between **2 and 5** gives $\{A, B\}$ and $\{C, D\}$.

Problem 6. Calibrating a parcel scale

(a) From 6 to 10 kg the sample mean increases by $9.4 - 5.8 = 3.6$ kg. From 10 to 14 kg it increases by $14.0 - 9.4 = 4.6$ kg. The increase is larger on the **10–14 kg interval**.

(b) **(i)** At $x = 16$, the fitted mass is $0.4 + 0.9 \times 16 + 0.3 \times 6 = \mathbf{16.6}$ kg. **(ii)** The slope increases from 0.9 below 10 kg to $0.9 + 0.3 = 1.2$ above 10 kg. A 1-kg increase entirely above the knot predicts **1.2 kg** more mass. The 0.3 is the increase in slope at the knot.

(c) **(i)** Equal fold sizes give a CV MSE of $(0.4+0.5+0.6+0.7+0.8)/5 = 0.6$ for the one-knot spline. **(ii)** CV selects the **one-knot spline** ($0.6 < 0.9 < 1.1$). Training MSE selects the **more flexible spline** (0.3). Model selection favors candidates whose CV error happens to be small. Evaluate the chosen model on fresh parcels kept out of both selection and fitting.

(d) **(i)** Under the prediction-interval rule, **hold** the parcel because $14.0 > 13.5$. Under the proposed mean-confidence-interval rule, **release** it because $13.2 \leq 13.5$.

(ii) No. The confidence interval quantifies uncertainty in the mean certified mass at a given reading. Individual masses can exceed 13.5 kg even when that mean is below the limit. CV MSE estimates average squared prediction error across all parcels. The claimed bound concerns $P(Y > 13.5 \mid \text{released})$, which these summaries do not control.

(iii) The estimated overweight fraction among releases is $6/80 = \mathbf{7.5}\%$. **Neither conclusion is established.** Since $0.211 > 0.05$, the audit gives insufficient evidence that the population rate exceeds 5%. Failing to reject the manager's claim also leaves compliance unverified.

Problem 7. Bonus questions

(a) Write E_1, E_2 for the rejection events. Their union contains E_1 and has probability at most $P(E_1) + P(E_2)$, giving

$$\mathbf{0.05} \leq P(E_1 \cup E_2) \leq \mathbf{0.10}.$$

Identical rejection events attain the lower bound. Disjoint rejection events attain the upper bound.

(b) The rescaled points are $(\pm s, \pm 1)$. In the left/right partition the centers are $(\pm s, 0)$, so each point contributes 1 and the total is 4. In the upper/lower partition the centers are $(0, \pm 1)$, so the total is $4s^2$. Thus **upper/lower wins for** $0 < s < 1$, the partitions **tie at** $s = 1$, and **left/right wins for** $s > 1$. Standardization divides the horizontal and vertical coordinates by $2s/\sqrt{3}$ and $2/\sqrt{3}$, respectively. The four resulting points form a square for every $s > 0$, so **the two costs always tie**.