Home Assignment 2 Solutions

STAT 154/254: Modern Statistical Prediction & Machine Learning

Ajay Sharma — Fall 2024

Problem 1: Converting $\star$-values to p-values

Let X be a random variable on $\mathbb{R}$ with density $p(x)$. Assume $p(x) > 0$ for all $x \in \mathbb{R}$. Define

$$F_1(s) = P(X \leq s), \quad F_2(s) = P(X \geq s).$$

Show that

$$F_1(X) \sim \text{Unif}([0,1]) \quad \text{and} \quad F_2(X) \sim \text{Unif}([0,1]).$$

(Please avoid the confusing notation $P(X \leq X)$.)

Solution. Since we have $p(x) > 0$ for all $x \in \mathbb{R}$, the function

$$F_1(s) = \int_{-\infty}^{s} p(x)\, dx$$

is strictly increasing, and hence possesses an inverse F_1^{-1}. Thus, for any $s \in [0,1]$ we have

$$\mathbb{P}\big(F_1(X) \leq s\big) = \mathbb{P}\big(X \leq F_1^{-1}(s)\big) = F_1\big(F_1^{-1}(s)\big) = s.$$

This shows that $F_1(X)$ has a Uniform$[0,1]$ distribution. Similarly, since $p(x) > 0$ for all $x \in \mathbb{R}$, we have $F_2(s) = 1 - F_1(s)$ for all $s \in [0,1]$, hence by the above:

$$\mathbb{P}\big(F_2(X) \leq s\big) = \mathbb{P}\big(F_1(X) \geq 1 - s\big) = 1 - \mathbb{P}\big(F_1(X) \leq 1 - s\big) = 1 - (1 - s) = s.$$

This shows that $F_2(X)$ has a Uniform$[0,1]$ distribution as well.

Problem 2: Logistic regression log-likelihood

Let $(x_i, y_i)_{i=1}^n$ be iid samples with $x_i \in \mathbb{R}^d$ and label $y_i \in \{-1, 1\}$. Define

$$P_\beta\big(Y = 1 \mid X\big) = \frac{\exp(\langle X, \beta\rangle)}{1 + \exp(\langle X, \beta\rangle)}, \quad P_\beta\big(Y = -1 \mid X\big) = 1 - P_\beta\big(Y = 1 \mid X\big).$$

Let the log-likelihood be

$$\ell_n(\beta) = \sum_{i=1}^n \log P_\beta(Y = y_i \mid X = x_i).$$

Write down and simplify $\log \ell_n(\beta)$. Compute its gradient $\nabla_\beta[\log \ell_n(\beta)]$ and its Hessian $\nabla_\beta^2[\log \ell_n(\beta)]$.

Solution. By assumption we have

$$P_\beta(Y = +1 \mid X) = \frac{e^{X^\top \beta}}{1 + e^{X^\top \beta}} = \frac{1}{1 + e^{-X^\top \beta}}.$$

Therefore,

$$P_\beta(Y = -1 \mid X) = 1 - P_\beta(Y = +1 \mid X) = 1 - \frac{e^{X^\top \beta}}{1 + e^{X^\top \beta}} = \frac{1}{1 + e^{X^\top \beta}}.$$

To combine both expressions, we can write

$$P_\beta(Y \mid X) = \frac{1}{1 + e^{-Y X^\top \beta}}.$$

Then, using the fact that $P_\beta(X) = P(X)$ does not depend on β, we find the log-likelihood

$$\log \mathcal{L}_n(\beta) = \sum_{i=1}^n \log\Big(\frac{1}{1 + e^{-Y_i X_i^\top \beta}}\Big) + \sum_{i=1}^n \log P(X_i).$$

Now we take the gradient with respect to β. The second term vanishes since it is a constant, and to the first term we can apply the chain rule, yielding:

$$\nabla_\beta \log \mathcal{L}_n(\beta) = \nabla_\beta \sum_{i=1}^n \big[-\log\big(1+e^{-Y_i X_i^\top \beta}\big)\big] = \sum_{i=1}^n \frac{-\nabla_\beta e^{-Y_i X_i^\top \beta}}{1 + e^{-Y_i X_i^\top \beta}} = \sum_{i=1}^n \frac{Y_i X_i e^{-Y_i X_i^\top \beta}}{1 + e^{-Y_i X_i^\top \beta}} = \sum_{i=1}^n \frac{Y_i X_i}{1 + e^{Y_i X_i^\top \beta}}.$$

To compute the Hessian, we again use the chain rule:

$$\nabla_\beta^2 \log \mathcal{L}_n(\beta) = \sum_{i=1}^n \nabla_\beta\Big(\frac{Y_i X_i}{1 + e^{Y_i X_i^\top \beta}}\Big) = \sum_{i=1}^n Y_i X_i \frac{\nabla_\beta e^{Y_i X_i^\top \beta}}{\big(1 + e^{Y_i X_i^\top \beta}\big)^2} = \sum_{i=1}^n \frac{(Y_i)^2 X_i X_i^\top e^{Y_i X_i^\top \beta}}{\big(1 + e^{Y_i X_i^\top \beta}\big)^2}.$$

Notice that $(Y_i)^2 = 1$ whether $Y_i = 1$ or $Y_i = -1$. Also, we can write

$$\frac{e^{Y_i X_i^\top \beta}}{\big(1 + e^{Y_i X_i^\top \beta}\big)^2} = \frac{1}{1 + e^{Y_i X_i^\top \beta}} \cdot \frac{1}{1 + e^{-Y_i X_i^\top \beta}}.$$

Therefore,

$$\nabla_\beta^2 \log \mathcal{L}_n(\beta) = -\sum_{i=1}^n X_i X_i^\top \frac{1}{1 + e^{Y_i X_i^\top \beta}} \cdot \frac{1}{1 + e^{-Y_i X_i^\top \beta}}.$$

Problem 3: Projection Matrices I

Let $P_1, P_2 \in \mathbb{R}^{n\times n}$ be two projection matrices (i.e. $P_i^\top = P_i$ and $P_i^2 = P_i$) satisfying $P_1P_2 = 0$. Let $\operatorname{rank}(P_i) = r_i$ with $r_1 + r_2 \le n$. Define the diagonal matrices

$$D_1 = \operatorname{diag}(\underbrace{1,\dots,1}_{r_1},\underbrace{0,\dots,0}_{n-r_1}), \quad D_2 = \operatorname{diag}(\underbrace{1,\dots,1}_{r_2},\underbrace{0,\dots,0}_{n-r_2}).$$

Show that there exists an orthogonal $U \in \mathbb{R}^{n\times n}$ such that

$$P_1 = U\, D_1\, U^\top, \quad P_2 = U\, D_2\, U^\top,$$

i.e. P_1 and P_2 are simultaneously diagonalizable. One approach is:

i. Show there are orthogonal $V_1, V_2 \in \mathbb{R}^{n\times n}$ with $P_1 = V_1D_1V_1^\top$ and $P_2 = V_2D_2V_2^\top$.

ii. Let $\tilde{U}_1$ be the first r_1 columns of V_1, and let $\tilde{U}_2$ be columns $r_1+1,\dots,r_1+r_2$ of V_2. Show $P_1 = \tilde{U}_1\tilde{U}_1^\top$ and $P_2 = \tilde{U}_2\tilde{U}_2^\top$.

iii. Prove $\tilde{U}_1^\top\tilde{U}_2 = 0_{r_1\times r_2}$ using $P_1P_2 = 0$ and $\tilde{U}_i^\top\tilde{U}_i = I_{r_i}$.

iv. Extend $\{\tilde{U}_1, \tilde{U}_2\}$ to an orthonormal basis $U \in \mathbb{R}^{n\times n}$.

v. Conclude $P_1 = UD_1U^\top$ and $P_2 = UD_2U^\top$.

Solution.

i. Fix $i \in \{1,2\}$, and let us diagonalize $P_i = W_i\Sigma_iW_i^\top$, where Σ_i is diagonal and W_i is orthogonal. Now observe that $P_i^2 = P_i$ is equivalent to $W_i\Sigma_i^2W_i^\top = W_i\Sigma_iW_i^\top$, so, canceling the W_i matrix on either side, we get $\Sigma_i^2 = \Sigma_i$. This shows that every eigenvalue λ of P_i satisfies $\lambda^2 = \lambda$, hence λ must equal 0 or 1. In summary, all of the eigenvalues of P_i are 0 or 1.
Now we prove the result. Since all of the eigenvalues of P_i are either 0 or 1, there exists a permutation matrix Π_i such that $\Sigma_i = \Pi_iD_i\Pi_i^\top$. Thus,

$$P_1 = W_1\Sigma_1W_1^\top = W_1\Pi_1D_1\Pi_1^\top W_1^\top = (W_1\Pi_1)\, D_1\, (W_1\Pi_1)^\top.$$

Note that $V_1 = W_1\Pi_1$ is itself an orthogonal matrix, so we have shown $P_1 = V_1D_1V_1^\top$, as desired. The same proof applies for $i = 2$, since we can get a permutation matrix Π_2 such that $\Sigma_2 = \Pi_2D_2\Pi_2^\top$.

ii. For $i \in \{1,2\}$, write v_j^i for the jth column of V_i, and recall that we can write the diagonalization $P_i = V_iD_iV_i^\top$ as

$$P_i = \sum_{j=1}^{n}(D_i)_{jj}\, v_j^i\, (v_j^i)^\top.$$

Since each D_i has only 0 or 1 on its diagonal, we can simplify the sum. Indeed, we get

$$P_1 = \sum_{j=1}^{r_1} v_j^1\, (v_j^1)^\top = \tilde{U}_1\tilde{U}_1^\top \quad \text{and} \quad P_2 = \sum_{j=r_1+1}^{r_1+r_2} v_j^2\, (v_j^2)^\top = \tilde{U}_2\tilde{U}_2^\top$$

as claimed.

iii. By the above, we have $P_1P_2 = 0$ hence $\tilde{U}_1\tilde{U}_1^\top\,\tilde{U}_2\tilde{U}_2^\top = 0$. Now multiply on the left by $\tilde{U}_1^\top$ and on the right by $\tilde{U}_2$ to get

$$0 = \tilde{U}_1^\top\, 0\, \tilde{U}_2 = \tilde{U}_1^\top\tilde{U}_1\,\tilde{U}_1^\top\tilde{U}_2\,\tilde{U}_2^\top\tilde{U}_2 = I\,\tilde{U}_1^\top\tilde{U}_2 = \tilde{U}_1^\top\tilde{U}_2.$$

iv. Notice that $\tilde{U}_1^\top\tilde{U}_2 = 0$ means that every column of $\tilde{U}_1$ is orthogonal to every column of $\tilde{U}_2$. Thus, combining this with $\operatorname{rank}(\tilde{U}_i) = r_i$ for $i \in \{1,2\}$, we find that the matrix $\tilde{U} := [\tilde{U}_1, \tilde{U}_2]$ satisfies $\operatorname{rank}(\tilde{U}) = r_1+r_2$. Now we can let $\tilde{U}_3$ be any matrix whose columns are an orthonormal basis for the orthogonal complement of $\operatorname{col}(\tilde{U})$. It follows by construction that $\bar{U} := [\tilde{U}_1, \tilde{U}_2, \tilde{U}_3]$ is orthogonal.

v. This follows from the construction, since, for each $i \in \{1, 2\}$, the columns of $\bar{U}$ are exactly the columns of $\tilde{U}_i$ at the indices for which $\bar{D}_i$ is non-zero, in which case they are equal to the corresponding columns of V_i. More explicitly, if we write u_j for the jth column of $\bar{U}$, then we can just compute:

$$P_1 = \sum_{j=1}^{r_1} v_j^1 (v_j^1)^\top = \sum_{j=1}^{r_1} u_j u_j^\top = \sum_{j=1}^{n} (D_1)_{jj}\, u_j u_j^\top = \bar{U} D_1 \bar{U}^\top,$$

and

$$P_2 = \sum_{j=r_1+1}^{r_1+r_2} v_j^2 (v_j^2)^\top = \sum_{j=r_1+1}^{n} u_j u_j^\top = \sum_{j=1}^{n} (D_2)_{jj}\, u_j u_j^\top = \bar{U} D_2 \bar{U}^\top.$$

Problem 4: Projection Matrices II

Let $X \in \mathbb{R}^{n\times d}$ with $n \geq d$ and $\operatorname{rank}(X) = d$. Define

$$P = X\big(X^\top X\big)^{-1}X^\top \in \mathbb{R}^{n\times n}.$$

For any subset $T \subseteq \{1, 2, \ldots, d\}$ of size $|T| = t$, let X_T be the $n \times t$ submatrix of X with columns indexed by T, and set

$$P_T = X_T\big(X_T^\top X_T\big)^{-1}X_T^\top, \quad P_1 = I_n - P, \quad P_2 = P - P_T.$$

Show that:

i. $P = P^\top$, $P_T = P_T^\top$, $P^2 = P$, $P_T^2 = P_T$, so P and P_T are projections.
ii. $P_1 = P_1^\top$, $P_1^2 = P_1$, so P_1 is a projection of rank $n - d$.
iii. $PX = X$, hence $PX_T = X_T$.
iv. $PP_T = P_TP = P_T$.
v. $P_2 = P_2^\top$, $P_2^2 = P_2$, so P_2 is a projection of rank $d - t$.
vi. $P_1P_2 = 0$.

Solution.

i. We can easily check $P^2 = P$ by using matrix associativity and the definition of the inverse:

$$P^2 = X\,(X^\top X)^{-1}X^\top\,X\,(X^\top X)^{-1}X^\top = X\,(X^\top X)^{-1}\,(X^\top X)\,(X^\top X)^{-1}X^\top = X\,(X^\top X)^{-1}X^\top = P.$$

We can check $P^\top = P$ by using properties of the transpose:

$$P^\top = \big(X\,(X^\top X)^{-1}X^\top\big)^\top = X^\top\big((X^\top X)^{-1}\big)^\top X = X^\top\big((X^\top X)^\top\big)^{-1}X = X^\top(X^\top X)^{-1}X = P.$$

The properties $P_T^2 = P_T$ and $P_T^\top = P_T$ are proved exactly the same.

ii. Using the properties in the previous part, we compute

$$P_1^\top = (I - P)^\top = I^\top - P^\top = I - P = P_1, \quad P_1^2 = (I - P)^2 = I - 2P + P^2 = I - 2P + P = I - P = P_1.$$

Since P is a projection matrix with $\operatorname{rank}(P) = d$, it follows that P_1 is a projection matrix with $\operatorname{rank}(P_1) = n - d$.

iii. We can simply check:

$$PX = X\,(X^\top X)^{-1}X^\top\,X = X\,(X^\top X)^{-1}\,(X^\top X) = X.$$

Since X_T is a matrix consisting of a subset of the columns of X, we see that $PX = X$ implies $PX_T = X_T$.

iv. Using the previous part, we get

$$PP_T = PX_T\,(X_T^\top X_T)^{-1}X_T^\top = X_T\,(X_T^\top X_T)^{-1}X_T^\top = P_T.$$

So, using the fact that both P and P_T are symmetric, we get

$$P_TP = P_T^\top P^\top = (PP_T)^\top = P_T^\top = P_T.$$

v. Using the previous part, we get

$$P_2^2 = (P - P_T)^2 = P^2 - PP_T - P_TP + P_T^2 = P - P_T - P_T + P_T = P - P_T = P_2.$$

Since P is a projection matrix with $\operatorname{rank}(P) = d$ and P_T is a projection matrix with $\operatorname{rank}(P_T) = t$, it follows that P_2 is a projection matrix with $\operatorname{rank}(P_2) = d - t$.

vi. Putting all the pieces together, we get:

$$P_1P_2 = (I - P)\,(P - P_T) = P - P_T - P^2 + PP_T = P - P_T - P + P_T = 0.$$

Problem 5: F-statistic follows the F distribution

Let $(x_i, y_i)_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$. Write

$$X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix} \in \mathbb{R}^{n\times d}, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n,$$

and assume $n \geq d$ and $\operatorname{rank}(X) = d$. Let $S \subseteq \{1, \ldots, d\}$ and S^c its complement, with $|S^c| = d_0$. Denote by X_{S^c} the $n \times d_0$ submatrix of X with columns in S^c. Under the null hypothesis $y_i = \langle \beta_{S^c}, x_{i,S^c}\rangle + \varepsilon_i$, $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, define

$$\mathrm{RSS}_1 = \min_{\beta\in\mathbb{R}^d} \|y - X\beta\|_2^2, \quad \mathrm{RSS}_0 = \min_{\beta\in\mathbb{R}^{d_0}} \left\|y - X_{S^c}\beta\right\|_2^2.$$

One shows that

$$\mathrm{RSS}_0 - \mathrm{RSS}_1 \sim \sigma^2\chi^2(d - d_0), \quad \mathrm{RSS}_1 \sim \sigma^2\chi^2(n - d),$$

and that $\mathrm{RSS}_0 - \mathrm{RSS}_1$ is independent of RSS_1. Hence the statistic

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(d - d_0)}{\mathrm{RSS}_1/(n - d)}$$

follows an $F_{d-d_0,\,n-d}$ distribution.

Outline of proof:

i. Show $\mathrm{RSS}_1 = \varepsilon^\top(I_n - X(X^\top X)^{-1}X^\top)\,\varepsilon$ and $\mathrm{RSS}_0 = \varepsilon^\top(I_n - X_{S^c}(X_{S^c}^\top X_{S^c})^{-1}X_{S^c}^\top)\,\varepsilon$, where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$.

ii. Define projection matrices

$$P_1 = I_n - X(X^\top X)^{-1}X^\top, \quad P_2 = X(X^\top X)^{-1}X^\top - X_{S^c}(X_{S^c}^\top X_{S^c})^{-1}X_{S^c}^\top,$$

so that $\mathrm{RSS}_1 = \varepsilon^\top P_1\varepsilon$, $\mathrm{RSS}_0 - \mathrm{RSS}_1 = \varepsilon^\top P_2\varepsilon$. Use Q4 to show P_1 and P_2 are projections of ranks $n - d$ and $d - d_0$, with $P_1P_2 = 0$.

iii. By Q3, there is orthogonal U with $P_1 = UD_1U^\top$, $P_2 = UD_2U^\top$. Conclude $\varepsilon^\top P_1\varepsilon \sim \sigma^2\chi^2(n-d)$, $\varepsilon^\top P_2\varepsilon \sim \sigma^2\chi^2(d - d_0)$, and they are independent.

Solution.

i. By lecture, $\beta' = (X^\top X)^{-1}X^\top y$. So we can write

$$\begin{aligned}
\mathrm{RSS}_1 &= \min_{\beta\in\mathbb{R}^d} \|y - X\beta\|_2^2 \\
&= \left\|\, y - X(X^\top X)^{-1}X^\top y\right\|_2^2 \\
&= \left\|\, (I - X(X^\top X)^{-1}X^\top)\, y\right\|_2^2 \\
&= \left\|\, (I - X(X^\top X)^{-1}X^\top)\, \varepsilon\right\|_2^2 \\
&= \varepsilon^\top\, (I - X(X^\top X)^{-1}X^\top)\, \varepsilon.
\end{aligned}$$

The same steps (with $X \to X_{S^c}$) give RSS_0.

ii. Simply take

$$P = X\,(X^\top X)^{-1}X^\top, \quad T = S^c.$$

iii. Using $P_i = U\,D_i\,U^\top$ from Q3 and $z = U^\top\varepsilon \sim N(0, \sigma^2 I)$: $\varepsilon^\top P_i\varepsilon = \varepsilon^\top\,U\,D_i\,U^\top\varepsilon = z^\top D_i z$. Hence
   (a) $\varepsilon^\top P_1\varepsilon = z^\top D_1 z = \sum_{j=1}^{n-d} z_j^2$, so $\varepsilon^\top P_1\varepsilon \sim \sigma^2\chi^2(n - d)$.
   (b) Since the two chi-square sums involve disjoint subsets of the independent z_j, they are independent.