Home Assignment 3 Solutions

STAT 154/254: Modern Statistical Prediction & Machine Learning

Ajay Sharma — Fall 2024

Problem 1: Bayes prediction rule for LDA

Let the distribution of (X, Y) be $\mathbb{P}(Y = 1) = \pi_1$, $\mathbb{P}(Y = 0) = 1 - \pi_1$, and $\big[X \mid Y = k\big] \sim \mathcal{N}(\mu_k, \Sigma)$ for $k = 0, 1$. Here $\pi_1 \in [0, 1]$, $\mu_1, \mu_0 \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$. Please find the expression $L_k(x)$ for $k = 0, 1$, which are linear functions in x (L_k can also depend on other parameters such as π_1, μ_k, Σ) such that the decision rule

$$\hat{k}(x) = \arg\max_k \mathbb{P}(Y = k \mid X = x)$$

can be rewritten as

$$\hat{k}(x) = \arg\max_k L_k(x).$$

Solution. To begin, we use Bayes' rule to get

$$\mathbb{P}(Y = k \mid X = x) = \mathbb{P}(X = x \mid Y = k)\,\frac{\mathbb{P}(Y = k)}{\mathbb{P}(X = x)},$$

and we note that the denominator does not depend on k. Thus, we have

$$\arg\max_k \mathbb{P}(Y = k \mid X = x) = \arg\max_k \mathbb{P}(X = x \mid Y = k)\,\mathbb{P}(Y = k),$$

and we can compute the right side. The definition of LDA is exactly that $\mathbb{P}(Y = k) = \pi_k$ and

$$\mathbb{P}(X = x \mid Y = k) = (2\pi)^{-\frac{d}{2}}(\det\Sigma)^{-\frac{1}{2}}\exp\Big(-\tfrac{1}{2}(x-\mu_k)^\top\Sigma^{-1}(x-\mu_k)\Big).$$

By taking the logarithm and ignoring the terms that do not depend on k, we get:

$$\begin{aligned}\arg\max_k \mathbb{P}(Y = k \mid X = x) &= \arg\max_k\Big(-\tfrac{1}{2}(x-\mu_k)^\top\Sigma^{-1}(x-\mu_k) + \log\pi_k\Big)\\ &= \arg\max_k\Big(x^\top\Sigma^{-1}\mu_k - \tfrac{1}{2}\mu_k^\top\Sigma^{-1}\mu_k + \log\pi_k\Big).\end{aligned}$$

Therefore, we find that the linear functions $f_k : \mathbb{R}^d \to \mathbb{R}$ defined via

$$f_k(x) = x^\top\Sigma^{-1}\mu_k - \tfrac{1}{2}\mu_k^\top\Sigma^{-1}\mu_k + \log\pi_k$$

are as desired.

Problem 2: Maximum likelihood estimator of mean and covariance in LDA

Let $(x_i, y_i)_{i=1}^n \subset \mathbb{R}^d \times \{0,1\}$ be i.i.d. samples from the same (X,Y) as in Q1. Please write down and simplify the log-likelihood function

$$\log \mathcal{L}(\pi_1, \mu_1, \mu_0, \Sigma \mid (x_i, y_i)_{i=1}^n) = \log \left[\prod_{i=1}^n \big(p_{X|Y}(x_i \mid y_i) \, \mathbb{P}(Y = y_i) \big) \right].$$

Let

$$(\hat{\pi}_1, \hat{\mu}_1, \hat{\mu}_0, \hat{\Sigma}) = \arg \max_{\pi_1, \mu_1, \mu_0, \Sigma} \log \mathcal{L}(\pi_1, \mu_1, \mu_0, \Sigma \mid (x_i, y_i)_{i=1}^n).$$

Please give explicit expressions for $\hat{\pi}_1, \hat{\mu}_1, \hat{\mu}_0, \hat{\Sigma}$.

Solution. The log-likelihood of the model parameters is exactly:

$$\begin{aligned}
\log \mathcal{L}(\pi_1, \mu_1, \mu_0, \Sigma \mid (x_i, y_i)_{i=1}^n) = &-\frac{nd}{2} \log(2\pi) - \frac{n}{2} \log\big(\det(\Sigma)\big) \\
&+ \sum_{i=1}^n \mathbf{1}_{\{y_i = 0\}} \Big(-\tfrac{1}{2} (x_i - \mu_0)^\top \Sigma^{-1} (x_i - \mu_0) + \log(1 - \pi_1) \Big) \\
&+ \sum_{i=1}^n \mathbf{1}_{\{y_i = 1\}} \Big(-\tfrac{1}{2} (x_i - \mu_1)^\top \Sigma^{-1} (x_i - \mu_1) + \log \pi_1 \Big),
\end{aligned}$$

so we can differentiate this and set the result equal to zero in order to find the MLE parameters. In order to simplify notation, let us also write

$$N_0 := \#\{1 \le i \le n : y_i = 0\}, \quad N_1 := \#\{1 \le i \le n : y_i = 1\}.$$

First let's take the derivative with respect to π_1:

$$\frac{\partial \mathcal{L}}{\partial \pi_1} = -\frac{N_0}{1 - \pi_1} + \frac{N_1}{\pi_1}.$$

Setting this equal to 0 and using the fact that $N_0 + N_1 = n$ yields exactly

$$\hat{\pi}_1 = \frac{N_1}{n}.$$

(Analogously we can also get $\hat{\pi}_0 = N_0/n$.)
Second, we take the derivative with respect to μ_0:

$$\frac{\partial \mathcal{L}}{\partial \mu_0} = \sum_{i=1}^n \mathbf{1}_{\{y_i = 0\}} \Sigma^{-1} (x_i - \mu_0).$$

By multiplying by Σ and canceling, we find that $\partial \mathcal{L} / \partial \mu_0 = 0$ is equivalent to $\sum_{i=1, y_i = 0}^n x_i = \sum_{i=1, y_i=0}^n \mu_0$, hence the MLE is just

$$\hat{\mu}_0 = \frac{1}{N_0} \sum_{i=1 \,:\, y_i = 0}^n x_i.$$

The same line of reasoning shows that

$$\hat{\mu}_1 = \frac{1}{N_1} \sum_{i=1 \,:\, y_i = 1}^n x_i.$$

Finally, we take the derivative with respect to Σ, for which we will need the fact that $\frac{d}{d\Sigma}\log(\det(\Sigma)) = \Sigma^{-1}$. Then, we get:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \Sigma} = -\frac{n}{2}\Sigma^{-1} &+ \frac{1}{2}\sum_{i=1}^{n}\mathbf{1}_{\{y_i=0\}}\Sigma^{-1}(x_i-\mu_0)(x_i-\mu_0)^\top\Sigma^{-1} \\ &+ \frac{1}{2}\sum_{i=1}^{n}\mathbf{1}_{\{y_i=1\}}\Sigma^{-1}(x_i-\mu_1)(x_i-\mu_1)^\top\Sigma^{-1}.\end{aligned}$$

By setting this equal to zero, multiplying on the left and right by Σ, and rearranging:

$$\hat{\Sigma} = \frac{1}{n}\Big(\sum_{i=1}^{n}\mathbf{1}_{\{y_i=0\}}(x_i-\mu_0)(x_i-\mu_0)^\top + \sum_{i=1}^{n}\mathbf{1}_{\{y_i=1\}}(x_i-\mu_1)(x_i-\mu_1)^\top\Big).$$

Therefore, plugging in the MLE for μ_0 and μ_1 above, we conclude:

$$\hat{\Sigma} = \frac{1}{n}\Big(\sum_{i=1}^{n}\mathbf{1}_{\{y_i=0\}}(x_i-\hat{\mu}_0)(x_i-\hat{\mu}_0)^\top + \sum_{i=1}^{n}\mathbf{1}_{\{y_i=1\}}(x_i-\hat{\mu}_1)(x_i-\hat{\mu}_1)^\top\Big).$$

Note that MLE parameters can be interpreted very intuitively: $\hat{\pi}_0$ and $\hat{\pi}_1$ are exactly the proportion of labels within the given class; $\hat{\mu}_0$ and $\hat{\mu}_1$ are exactly the empirical means within each class; $\hat{\Sigma}$ is the pooled covariance matrix, where we normalize each data point by its class mean.

Problem 3: Property of AUC

Let $(x_i, y_i)_{i=1}^n \subset \mathbb{R}^d \times \{0,1\}$ be an i.i.d. validation dataset and let $\hat{f} : \mathbb{R}^d \to [0,1]$ be a continuous score function. Define the population true- and false-positive rates

$$\mathrm{TPR}_*(\hat{f}, t) = \mathbb{P}\big(\hat{f}(X) \geq t \mid Y = 1\big), \quad \mathrm{FPR}_*(\hat{f}, t) = \mathbb{P}\big(\hat{f}(X) \geq t \mid Y = 0\big).$$

As t varies over $[0,1]$, the ROC_* curve is the graph in the $\mathrm{FPR}_*, \mathrm{TPR}_*$ plane, and the area under it is $\mathrm{AUC}_*(\hat{f})$. Show that

$$\mathrm{AUC}_*(\hat{f}) \;=\; 1 - \mathrm{AUC}_*(1 - \hat{f}),$$

under the assumption $\mathbb{P}\big(\hat{f}(X) = t \mid Y = k\big) = 0$ for all $k \in \{0,1\}$ and $t \in [0,1]$.

Solution. By definition we have

$$\mathrm{AUC}_*(\hat{f}) = \int_0^1 \mathrm{TPR}_*\big(\mathrm{FPR}_*^{-1}(z)\big)\, dz,$$

where we have defined

$$\mathrm{TPR}_{\hat{f}}(t) := \mathbb{P}\big(\hat{f}(X) \geq t \mid Y = 1\big), \quad \mathrm{FPR}_{\hat{f}}(t) := \mathbb{P}\big(\hat{f}(X) \geq t \mid Y = 0\big).$$

In order to analyze $\mathrm{AUC}_*(1 - \hat{f})$, we calculate, for any $t \in [0,1]$:

$$\begin{aligned}
\mathrm{TPR}_{1-\hat{f}}(t) &= \mathbb{P}(1 - \hat{f}(X) \geq t \mid Y = 1) \\
&= \mathbb{P}(\hat{f}(X) \leq 1 - t \mid Y = 1) \\
&= \mathbb{P}(\hat{f}(X) < 1 - t \mid Y = 1) \\
&= 1 - \mathbb{P}(\hat{f}(X) \geq 1 - t \mid Y = 1) \\
&= 1 - \mathrm{TPR}_{\hat{f}}(1 - t),
\end{aligned}$$

and similarly

$$\mathrm{FPR}_{1-\hat{f}}(t) = 1 - \mathrm{FPR}_{\hat{f}}(1 - t).$$

From the latter, we get

$$\mathrm{FPR}_{1-\hat{f}}^{-1}(z) = 1 - \mathrm{FPR}_{\hat{f}}^{-1}(1 - z), \quad z \in [0,1].$$

Therefore,

$$\begin{aligned}
\mathrm{AUC}_*(1 - \hat{f}) &= \int_0^1 \mathrm{TPR}_{1-\hat{f}}\big(\mathrm{FPR}_{1-\hat{f}}^{-1}(z)\big)\, dz \\
&= \int_0^1 \mathrm{TPR}_{1-\hat{f}}\big(1 - \mathrm{FPR}_{\hat{f}}^{-1}(1 - z)\big)\, dz \\
&= \int_0^1 \big(1 - \mathrm{TPR}_{\hat{f}}(\mathrm{FPR}_{\hat{f}}^{-1}(1 - z))\big)\, dz \\
&= \int_0^1 \big(1 - \mathrm{TPR}_{\hat{f}}(\mathrm{FPR}_{\hat{f}}^{-1}(z'))\big)\, dz' \\
&= 1 - \int_0^1 \mathrm{TPR}_{\hat{f}}(\mathrm{FPR}_{\hat{f}}^{-1}(z'))\, dz' \\
&= 1 - \mathrm{AUC}_*(\hat{f}).
\end{aligned}$$

Problem 4: The MAP estimator of Bayes Generalized Linear Model (GLM)

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

$$\mathbb{P}(y_i = 1 \mid x_i) = \frac{\exp\langle x_i, \beta\rangle}{1 + \exp\langle x_i, \beta\rangle}, \quad \mathbb{P}(y_i = -1 \mid x_i) = 1 - \mathbb{P}(y_i = 1 \mid x_i).$$

The distribution of $(x_i)_{i\in[n]}$ is fixed and is irrelevant to this question. Further assume that $(x_i, y_i)_{i\in[n]}$ are mutually independent. We further assume a prior over β:

$$\Pi(\beta) = \frac{1}{Z}\exp\bigl(-\|\beta\|_1/\sigma_0\bigr), \quad Z = \int_{\mathbb{R}^d} \exp\bigl(-\|\beta\|_1/\sigma_0\bigr)\, d\beta,$$

where $\sigma_0 > 0$. Define the MAP estimator

$$\hat{\beta}_{\mathrm{MAP}} = \arg\max_\beta \mathbb{P}\bigl(\beta \mid (x_i, y_i)_{i=1}^n\bigr).$$

Requirement of the result: Please find a function $E(\beta)$ such that the following holds

$$\hat{\beta}_{\mathrm{MAP}} = \arg\min_\beta E(\beta),$$

and $E(\beta)$ is the summation of the negative log-likelihood function and a regularization term. (Hint: the final result does not involve the constant Z, so don't worry too much about the integration in Z. But you should explain why the final result does not involve the constant Z.)

Solution. By Bayes' rule, we have:

$$\mathbb{P}\bigl(\beta \mid (x_1, y_1), \ldots, (x_n, y_n)\bigr) = \frac{\mathbb{P}(y_1, \ldots, y_n \mid \beta, x_1, \ldots, x_n)\, \mathbb{P}(\beta)}{\mathbb{P}(y_1, \ldots, y_n \mid x_1, \ldots, x_n)}.$$

Since the bottom factor does not depend on β, it can be ignored from the maximization problem. Also, since $z \mapsto \log z$ is monotone, the maximization is equivalent if we instead compute the log of the probabilities. Thus, using the definitions, we have

$$\begin{aligned}
\arg\max_\beta \mathbb{P}\bigl(\beta \mid (x_i, y_i)_{i=1}^n\bigr) &= \arg\max_\beta \mathbb{P}(y_1, \ldots, y_n \mid \beta, x_1, \ldots, x_n)\, \mathbb{P}(\beta) \\
&= \arg\max_\beta \Bigl(\sum_{i=1}^n \log \frac{1}{1 + e^{-y_i x_i^\top \beta}} - \frac{\|\beta\|_1}{\sigma_0} - \log Z\Bigr).
\end{aligned}$$

Finally, note that the value Z depends on σ_0 but not β, so it can also be ignored from the optimization. Therefore, we have shown that

$$\arg\max_\beta \mathbb{P}\bigl(\beta \mid (x_i, y_i)\bigr) = \arg\min_\beta E(\beta), \quad E(\beta) = -\sum_{i=1}^n \log \frac{1}{1 + e^{-y_i x_i^\top \beta}} + \frac{\|\beta\|_1}{\sigma_0}.$$

Notice that E is exactly the negative log-likelihood of logistic regression, plus an ℓ_1 regularization term which encourages sparsity.