Practical Session 04

Linear Classification

1 Naive Bayes

Problem 1: In LDA we assume that all classes share the same covariance matrix $\boldsymbol{\Sigma}$. The Naive Bayes classifier assumes that all d features of a sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$ are conditionally independent given the class, i.e.

$$p(x_1, x_2, \ldots, x_d|y) = \prod_{i=1}^{d} p(x_i|y)$$

In the case of continuous data where the likelihood is a normal distribution, this corresponds to **diagonal** covariance matrices that however are different for each class (not shared). The generative process is thus:

$$p(\boldsymbol{x}|y = c) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

Show that using different $\boldsymbol{\Sigma}_c$'s for each class leads to quadratic decision boundaries.

For simplicity, let's consider a binary classification problem $\mathcal{C} = \{0, 1\}$. Decision boundary DB can be defined as a set of points, for which the probabilities of both classes are equal. More formally we can write this as

$$DB = \{\boldsymbol{x} : p(y = 1 \mid \boldsymbol{x}) = p(y = 0 \mid \boldsymbol{x})\}$$

Assuming that $p(y = c \mid \boldsymbol{x}) \neq 0 \quad \forall \boldsymbol{x}, c$

$$\begin{aligned} DB &= \{\boldsymbol{x} : \frac{p(y = 1 \mid \boldsymbol{x})}{p(y = 0 \mid \boldsymbol{x})} = 1\} \\ &= \{\boldsymbol{x} : \ln \frac{p(y = 1 \mid \boldsymbol{x})}{p(y = 0 \mid \boldsymbol{x})} = 0\} \end{aligned}$$

This means, that finding the decision boundary is equivalent to solving $\ln \frac{p(y=1|\boldsymbol{x})}{p(y=0|\boldsymbol{x})} = 0$ for $\boldsymbol{x}$.

For Gaussian discriminant analysis with $p(\boldsymbol{x} \mid y = c) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$

$$
\begin{aligned}
\ln \frac{p(y = 1 \mid \boldsymbol{x})}{p(y = 0 \mid \boldsymbol{x})} &= \ln \frac{p(\boldsymbol{x} \mid y = 1)p(y = 1)p(\boldsymbol{x})}{p(\boldsymbol{x})p(\boldsymbol{x} \mid y = 0)p(y = 0)} \\
&= \ln\left(p(\boldsymbol{x} \mid y = 1)p(y = 1)\right) - \ln\left(p(\boldsymbol{x} \mid y = 0)p(y = 0)\right) \\
&= \ln \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) - \ln \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) + \ln \frac{\pi_1}{\pi_0} \\
&= -\frac{1}{2} \ln(2\pi)^D |\boldsymbol{\Sigma}_1| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1) \\
&\quad + \frac{1}{2} \ln(2\pi)^D |\boldsymbol{\Sigma}_0| + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_0) + \ln \frac{\pi_1}{\pi_0} \\
&= -\frac{1}{2}\boldsymbol{x}^T \boldsymbol{\Sigma}_1^{-1} \boldsymbol{x} + \boldsymbol{x}^T \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 - \frac{1}{2}\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 \\
&\quad + \frac{1}{2}\boldsymbol{x}^T \boldsymbol{\Sigma}_0^{-1} \boldsymbol{x} - \boldsymbol{x}^T \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \frac{1}{2}\boldsymbol{\mu}_0^T \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \frac{1}{2} \ln \frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|} + \ln \frac{\pi_1}{\pi_0} \\
&= \frac{1}{2}\boldsymbol{x}^T [\boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}_1^{-1}] \boldsymbol{x} + \boldsymbol{x}^T [\boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0] \\
&\quad - \frac{1}{2}\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_0^T \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \log \frac{\pi_1}{\pi_0} + \frac{1}{2} \ln \frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|} \\
&= \boldsymbol{x}^T \boldsymbol{W}_2 \boldsymbol{x} + \boldsymbol{w}_1^T \boldsymbol{x} + w_0
\end{aligned}
$$

Where

$$
\begin{aligned}
\boldsymbol{W}_2 &= \frac{1}{2}[\boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}_1^{-1}] \\
\boldsymbol{w}_1 &= \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \\
w_0 &= -\frac{1}{2}\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_0^T \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \log \frac{\pi_1}{\pi_0} + \frac{1}{2} \ln \frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|}
\end{aligned}
$$

If both classes had the same covariance matrix ($\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$), the quadratic terms $\frac{1}{2}\boldsymbol{x}^T \boldsymbol{\Sigma}_0^{-1} \boldsymbol{x}$ and $-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{\Sigma}_1^{-1} \boldsymbol{x}$ would cancel out and we would receive the linear decision boundary, like we did in the lecture (also, $\ln \frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|} = 0$). Otherwise, if $\boldsymbol{\Sigma}_0 \neq \boldsymbol{\Sigma}_1$ we get a quadratic decision boundary.

Derivation for the multiclass $C > 2$ setting is analogous, but more messy, so we leave it out.

2 Multi-Class Classification

Problem 2: Consider a generative classification model for C classes defined by prior class probabilities $p(y = c) = \pi_c$ and general class-conditional densities $p(\boldsymbol{x}|y = c, \boldsymbol{\theta}_c)$ where $\boldsymbol{x}$ is the input feature vector and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_c\}_{c=1}^C$ are further model parameters. Suppose we are given a training set $\mathcal{D} = \{(\boldsymbol{x}^{(n)}, y^{(n)})\}_{n=1}^N$ where $y^{(n)}$ is a binary target vector of length C that uses the 1-of-C(one-hot) encoding scheme, so that it has components $y_c^{(n)} = \delta_{ck}$ if pattern n is from class $y = k$. Assuming that the data points are iid, show that the maximum-likelihood solution for the prior probabilities is given by

$$\pi_c = \frac{N_c}{N}$$

where N_c is the number of data points assigned to class $y = c$.

The likelihood function of the parameters $\{\pi_c, \boldsymbol{\theta}_c\}_{c=1}^{C}$ is given by

$$p(\mathcal{D}|\{\pi_c, \boldsymbol{\theta}_c\}_{c=1}^{C}) = \prod_{n=1}^{N}\prod_{c=1}^{C}(p(\boldsymbol{x}^{(n)}|\boldsymbol{\theta}_c)\pi_c)^{y_c^{(n)}}$$

so the log-likelihood is given by

$$\log p(\mathcal{D}|\{\pi_c, \boldsymbol{\theta}_c\}_{c=1}^{C}) = \sum_{n=1}^{N}\sum_{c=1}^{C} y_c^{(n)} \log \pi_c + \text{const flin } \pi_c$$

In order to maximize the log likelihood with respect to π_c we need to preserve the constraint $\sum_c \pi_c = 1$. This can be done by introducing a Lagrange multiplier λ and maximizing

$$\sum_{n=1}^{N}\sum_{c=1}^{C} y_c^{(n)} \log \pi_c + \lambda\left(\sum_{c=1}^{C}\pi_c - 1\right).$$

Setting the derivative with respect to π_c equal to zero, we obtain

$$\lambda = \frac{1}{\pi_c}\sum_{n=1}^{N} y_c^{(n)} = \frac{1}{\pi_c}N_c.$$

Setting the derivative with respect to λ equal to zero, we obtain our constraint

$$\sum_{c=1}^{C}\pi_c = 1.$$

where we can now insert the previous result $\pi_c = \frac{N_c}{\lambda}$ and obtain

$$\lambda = \sum_c N_c = N.$$

So overall we obtain

$$\pi_c = \frac{N_c}{N}.$$

Problem 3: Using the same classification model as in the previous question, now suppose that the class-conditional densities are given by Gaussian distributions with a shared covariance matrix, so that

$$p(x|y = c, \boldsymbol{\theta}_c) = p(x|\boldsymbol{\theta}_c) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_c, \boldsymbol{\Sigma}).$$

Show that the maximum likelihood solution for the mean of the Gaussian distribution for class C_c is given by

$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{\{n|\boldsymbol{x}^{(n)} \in C_c\}} \boldsymbol{x}^{(n)}$$

which represents the mean of those feature vectors assigned to class C_c.

Similarly, show that the maximum likelihood solution for the shared covariance matrix is given by

$$\boldsymbol{\Sigma} = \sum_{c=1}^{C} \frac{N_c}{N} \mathbf{S}_c$$

where

$$\mathbf{S}_c = \frac{1}{N_c} \sum_{\{n|\boldsymbol{x}^{(n)} \in C_c\}} (\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)^T.$$

Thus $\boldsymbol{\Sigma}$ is given by a weighted average of the covariances of the data associated with each class, in which the weighting coefficients N_c/N are the prior probabilities of the classes.

If we substitute $p(\boldsymbol{x}|\boldsymbol{\theta}_c) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_c, \boldsymbol{\Sigma})$ into $\log p(\mathcal{D}|\{\pi_c, \boldsymbol{\theta}_c\}_{c=1}^C)$ and then use the definition of the multivariate Gaussian, we obtain

$$\log p(\mathcal{D}|\{\pi_c, \boldsymbol{\theta}_c\}_{c=1}^C) = \frac{-1}{2} \sum_{n=1}^{N} \sum_{c=1}^{C} y_c^{(n)} (\log |\boldsymbol{\Sigma}| + (\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c) + \log \pi_c)$$

Dropping terms independent of $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}$ we get

$$\log p(\mathcal{D}|\{\boldsymbol{\theta}_c\}_{c=1}^C) = \frac{-1}{2} \sum_{n=1}^{N} \sum_{c=1}^{C} y_c^{(n)} (\log |\boldsymbol{\Sigma}| + (\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c))$$

Setting the derivative of the above equation w.r.t $\boldsymbol{\mu}_c$, (obtained using $\frac{\partial}{\partial x}(x^T a) = \frac{\partial}{\partial x}(a^T x) = a$), to zero we get

$$\sum_{n=1}^{N} y_c^{(n)} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c) = 0.$$

Making use of what we learned in the last problem, i.e.

$$\sum_{n=1}^{N} y_c^{(n)} = N_c,$$

we can re-arrange this to obtain

$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{n=1}^{N} y_c^{(n)} \boldsymbol{x}^{(n)}.$$

Using the trace trick ($a = \text{Tr}(a)$ for $a \in \mathbb{R}$ and $\text{Tr}(\boldsymbol{ABC}) = \text{Tr}(\boldsymbol{BCA})$) we can rewrite our original expression $\log p(\mathcal{D}|\{\boldsymbol{\theta}_c\}_{c=1}^C) = \frac{-1}{2}\sum_{n=1}^N \sum_{c=1}^C y_c^{(n)}(\log|\boldsymbol{\Sigma}| + (\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c))$ as

$$\log p(\mathcal{D}|\{\boldsymbol{\theta}_c\}_{c=1}^C) = \frac{-1}{2}\sum_{n=1}^N \sum_{c=1}^C y_c^{(n)}(\log|\boldsymbol{\Sigma}| + \text{Tr}(\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)^T).$$

We can now use $\frac{\partial}{\partial \boldsymbol{A}} Tr(\boldsymbol{AB}) = \boldsymbol{B}^T$ and $\frac{\partial}{\partial \boldsymbol{A}} \ln|\boldsymbol{A}| = (\boldsymbol{A}^{-1})^T$ and $\ln|\boldsymbol{A}^{-1}| = -\ln|\boldsymbol{A}|$ to calculate the derivative w.r.t. $\boldsymbol{\Sigma}^{-1}$. Setting this to zero we obtain

$$\frac{1}{2}\sum_{n=1}^N \sum_{c=1}^C y_c^{(n)}(\boldsymbol{\Sigma} - (\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)^T) = 0.$$

Making use of $\sum_{n=1}^N y_c^{(n)} = N_c$, we can re-arrange this to obtain

$$\boldsymbol{\Sigma} = \sum_{c=1}^C \frac{N_c}{N}\mathbf{S}_c$$

,

where

$$\mathbf{S}_c = \frac{1}{N_c}\sum_{n=1}^N y_c^{(n)}(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)^T.$$

Note that we do not enforce that $\boldsymbol{\Sigma}$ should be symmetric, but the solution shows that it is automatically symmetric.

Problem 4: Error measures for classification

- ROC curve and AUC
- PR curve and AUC a.k.a. average precision

References

- https://en.wikipedia.org/wiki/Confusion_matrix
- https://en.wikipedia.org/wiki/Precision_and_recall
- https://en.wikipedia.org/wiki/Receiver_operating_characteristic
- Interactive demo for ROC curves - definitely do this in class http://www.navan.name/roc/
- PR AUC vs ROC AUC - https://stats.stackexchange.com/questions/7207/roc-vs-precision-and-recall-curves
- http://scikit-learn.org/stable/modules/model_evaluation.html