Lecturer: Prof. Homrighausen Scribe: Lei Yang

1 Classification set up

Suppose we have observations

$$\mathcal{D} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

Again, we want to estimate a function that maps X into Y that helps us predict as yet observed data, which is known as classifier.

let $Y \in \mathcal{G} = \{1, \ldots, G\}$

We again make predictions $\hat{Y}$ based on $\mathcal{D}$

Our loss function is now a $G \times G$ matrix L with zeros on the diagonals and $\ell(g, g')$ on the off diagonal $(g \neq g')$.

Again, we appeal to risk

$$R(\hat{g}) = \mathbb{E}_Z \ell_{\hat{g}}(Z)$$

If we use the law of total probability, this can be written

$$R(\hat{g}) = \mathbb{E}_X \sum_{y=1}^{G} \ell_{\hat{g}}(Z = (y, X)) \mathbb{P}(Y = y|X)$$

This can be minimized point wise over x, to produce

$$g^*(x) = \underset{g \in \mathcal{G}}{\operatorname{argmin}} \sum_{y=1}^{G} \ell_g(z = (y, x)) \mathbb{P}(Y = y|X = x)$$

This is the Bayes' classifier. Also, $R(g^*)$ is the Bayes' limit.

2 Best classifier

If we make specific choices for ℓ, we can find g^* exactly

As Y takes only a few values, zero-one prediction risk is natural

$$\ell_g(Z) = \mathbf{1}_{Y \neq g(X)}(Z) \Rightarrow R(g) = \mathbb{E}[\ell_g(Z)] = \mathbb{P}(g(X) \neq Y),$$

This means we want to label or classify a new observation (X, Y) such that $\hat{f}(X) = Y$ as often as possible.

Under this loss, we have

$$g^*(x) = \underset{g \in \mathcal{G}}{\operatorname{argmin}} \left[1 - \mathbb{P}(Y = g|X = x)\right] = \underset{g \in \mathcal{G}}{\operatorname{argmax}} \mathbb{P}(Y = g|X = x)$$

Suppose we encode a two-class response as $Y \in \{0, 1\}$

Let's continue to use squared error loss: $\ell_f(Z) = (Y - f(X))^2$

Then, the Bayes' rule is

$$m(X) = \mathbb{E}[Y|X] = \mathbb{P}(Y = 1|X)$$

Hence, we achieve the same Bayes' rule/limit with squared error classification by discretizing the probability:

$$g^*(x) = \mathbf{1}_{m(x)>1/2}(x)$$

3 Classification is easier than regression

Let $\hat{m}$ be any estimate of m

Let $\hat{g}(x) = \mathbf{1}_{\hat{m}(x)>1/2}(x)$

It can be shown that

$$\begin{aligned}
&\mathbb{P}(Y \neq \hat{g}(X)|X = x) - \mathbb{P}(Y \neq g^*(X)|X = x) = \\
&= (2m(x) - 1)(\mathbf{1}_{g^*(x)=1}(x) - \mathbf{1}_{\hat{g}(x)=1}(x)) \\
&= |2m(x) - 1|\mathbf{1}_{g^*(x)\neq\hat{g}(x)}(x) \\
&= 2|m(x) - 1/2|\mathbf{1}_{g^*(x)\neq\hat{g}(x)}(x)
\end{aligned}$$

Now

$$g^*(x) \neq \hat{g}(x) \Rightarrow |\hat{m}(x) - m(x)| \geq |m(x) - 1/2|$$

Therefore

$$\begin{aligned}
&\mathbb{P}(Y \neq \hat{g}(X)) - \mathbb{P}(Y \neq g^*(X)) = \\
&= \int (\mathbb{P}(Y \neq \hat{g}(X)|X = x) - \mathbb{P}(Y \neq g^*(X)|X = x))d\mathbb{P}_X(x) \\
&= \int 2|m(x) - 1/2|\mathbf{1}_{g^*(x)\neq\hat{g}(x)}(x)d\mathbb{P}_X(x) \\
&\leq 2\int |\hat{m}(x) - m(x)|\mathbf{1}_{g^*(x)\neq\hat{g}(x)}(x)d\mathbb{P}_X(x) \\
&\leq 2\int |\hat{m}(x) - m(x)|d\mathbb{P}_X(x)
\end{aligned}$$

If $\hat{m}$ gets close to m on average, we do good classifications. The converse is not true.

4 Bayes rule and class densities

Using Bayes' theorem

$$\begin{aligned} m(x) &= \mathbb{P}(Y = 1|X = x) \\ &= \frac{p(x|Y = 1)\mathbb{P}(Y = 1)}{\sum_{y\in\{0,1\}} p(x|Y = y)\mathbb{P}(Y = y)} \\ &= \frac{f_1(x)\pi}{f_1(x)\pi + f_0(x)(1-\pi)} \end{aligned}$$

We call $f_g(x)$ the class densities.

The Bayes' rule can be rewritten as

$$g^*(x) = \begin{cases} 1 & \text{if } \frac{f_1(x)}{f_0(x)} > \frac{1-\pi}{\pi} \\ 0 & \text{otherwise} \end{cases}$$

5 Find a classifier

All of these prior expressions for g^* give rise to classifiers

- Empirical risk minimization: Choose a set of classifiers Γ and find $\hat{g} \in \Gamma$ that minimizes some estimate of $R(g)$
- Regression: Find an estimate $\hat{m}$ and plug it in to the Bayes' rule
- Density estimation: Estimate f_g from the appropriate Z and $\hat{\pi} = \overline{Y}$ and plug them in to g^*

6 Linear classifiers

As our classifier $\hat{g}$ takes a discrete number of values, it is equivalent to partitioning the covariate space into regions. The boundaries between these regions are known as decision boundaries. A linear classifier is a $\hat{g}$ that produces linear decision boundaries

7 Bayes rule-ian approach

There are many techniques based on this idea

- Linear/quadratic discriminant analysis
  Estimates $p_g(x)$ assuming multivariate Gaussianity
- General nonparametric density estimators
- Naive Bayes Factors $p_g(x)$ assuming conditional independence

8 Discriminant analysis

Suppose that

$$p_g(x) \propto |\Sigma_g|^{-1/2} e^{-(x-\mu_g)^\top \Sigma_g^{-1}(x-\mu_g)/2}$$

Let's assume that $\Sigma_g \equiv \Sigma$.

Then the log-odds between two classes g, g' is:

$$\begin{aligned}\log \frac{\mathbb{P}(Y=g|X=x)}{\mathbb{P}(Y=g'|X=x)} &= \log \frac{p_g(x)}{p_{g'}(x)} + \log \frac{\pi_g}{\pi_{g'}} \\ &= \log \frac{\pi_g}{\pi_{g'}} - (\mu_g + \mu_{g'})^\top \Sigma^{-1}(\mu_g - \mu_{g'})/2 \\ &\quad + x^\top \Sigma^{-1}(\mu_g - \mu_{g'})\end{aligned}$$

This is linear in x, and hence has a linear decision boundary

The linear discriminant function is (proportional to) the log posterior:

$$\delta_g(x) = \log \pi_g + x^\top \Sigma^{-1}\mu_g - \mu_g^\top \Sigma^{-1}\mu_g/2$$

and we assign $g(x) = \text{argmin}_g\, \delta_g(x)$

This is just minimum Euclidean distance, weighted by the covariance matrix and prior probabilities

Now, we must estimate μ_g and Σ. If we use the intuitive estimators $\hat{\mu}_g = \overline{X}_g$ and $\hat{\Sigma} = \frac{1}{n-G}\sum_{g\in\mathcal{G}}\sum_{i\in g}(X_i - \hat{\mu}_g)(X_i - \hat{\mu}_g)^\top$ then we have produced linear discriminant analysis (LDA).

9 Quadratic discriminant analysis

If we drop the assumption regarding identical covariances, we get the following discriminant function:

$$\delta_g(x) = \log \pi_g + x^\top \Sigma_g^{-1}\mu_g - \mu_g^\top \Sigma_g^{-1}\mu_g/2 - \log|\Sigma_g|/2$$

where Σ_g can be estimated by the sample covariance of the observations in group g.

10 Reduced rank LDA

Part of the popularity of LDA is that it provides dimension reduction as well. The G class centroids μ_g must all lie in an affine subspace of dimension $G-1$ (presuming $G < p$). Let $\mathcal{H}_{G-1}$ be this subspace. If G is much less than p, this will be a substantial drop in dimension. In practice, we can compute LDA from spectral information:

$$\begin{aligned}\delta_g(x) &= \log \pi_g + x^\top \Sigma^{-1}\mu_g - \mu_g^\top \Sigma^{-1}\mu_g/2 \\ &\propto \log \pi_g + (x-\mu_g)^\top \Sigma^{-1}(x-\mu_g)/2\end{aligned}$$

So,

1. Spectrum: Form $\hat{\Sigma}_\lambda = UDU^\top$.

2. Sphere: Rewrite your data as $\tilde{X} \leftarrow D^{-1/2}U^\top X$.

3. Assign: Classify to the closest mean in transformed space, penalizing by estimate of prior probability.

We can ignore any information orthogonal to $\mathcal{H}_{G-1}$, as it contributes to each class equally (in the sphered space). So, project $\tilde{X}$ onto $\mathcal{H}_{G-1}$ and make distance computations there. When $G = 2, 3$, this means we can plot the projection onto $\mathcal{H}_{G-1}$ with no loss of information about the LDA solution.

If $G > 3$, then we may wish to project onto a reduced space $\mathcal{H}_L \subset \mathcal{H}_{G-1}$. We'd like $\mathcal{H}_L$ to maintain the most amount of information possible for assigning to classes.

This can be done via the following procedure:

1. Centroids: Compute $G \times p$ matrix M of class centroids.

2. Covariance: Form $\hat{\Sigma}$ as the common covariance matrix.

3. Sphere: $\tilde{M} = M\hat{\Sigma}^{-1/2}$.

4. Between Covariance: Find covariance matrix for $\tilde{M}$, call it B.

5. Spectrum Compute $B = VSV^\top$.

Now, $\text{span}(V_L) = \mathcal{H}_L$, where V_L denote the first L vectors in V. Also, the coordinates of the data in this space are $Z_k = v_k^\top \hat{\Sigma}^{-1/2} X$. These derived variables are commonly called canonical coordinates.