Classifier Design Based on Expected Risk (Loss/Cost) Minimization

Assume that our data samples $\{x_1, x_2, \dots\}$ are independent identically distributed (iid) instances of a real-valued n-dimensional random vector $\underline{X} \in \mathbb{R}^n$.

Let the probability density function (pdf) of $\underline{X}$ be a mixture (convex linear combination) of C class-conditional pdfs as follows:

$$p_{\underline{X}}(x) = p_{\underline{X}|L}(x|1)\,p_L(1) + \dots + p_{\underline{X}|L}(x|C)\,p_L(C)$$

Notation of Papoulis

$$= p(x|L=1)\,p(L=1) + \dots + p(x|L=C)\,p(L=C)$$

$$= \sum_{l=1}^{C} \underbrace{\overset{\text{CLASS-CONDITIONAL}}{p(x|L=l)}}_{\text{Data pdf for class label } l}\ \underbrace{\overset{\text{PRIOR}}{p(L=l)}}_{\text{Prior probability for class label } l}$$

Less precise but more convenient notation

Suppose that we want to design a decision rule (classifier) that selects the "best" choice among a discrete set of options for a given $x \in \mathbb{R}^n$: $D: \mathbb{R}^n \to \{1, \ldots, A\} \subset \mathbb{Z}$

for convenience assume options are indexed by integers.

In expected risk (loss/cost) minimization, the "best" option is to select the one which results in the smallest expected risk.

Let $\Lambda = \begin{array}{c|c} & j \\ \hline i & \begin{bmatrix} & \vdots & \\ \cdots & \lambda_{ij} & \cdots \\ & \vdots & \end{bmatrix} \end{array}$ be the loss matrix, such that λ_{ij} = "the risk/loss/cost associated with deciding on option i, given x comes from class label j"

Note that $i \in \{1, \ldots, A\}$ and $j \in \{1, \ldots, C\}$.

That is, in general, the set of decision options do <u>not</u> need to be the same as the set of class labels.

We choose $\lambda_{ij} \geq 0$ for all (i,j) pairs.

We want to find a decision rule D that minimizes expected risk/loss/cost. Overall expected risk is as follows:

$$\underset{X}{E}\left[Risk \right] = \int_{-\infty}^{\infty} Risk(\underset{D(x)}{Decide}|x)\, p_X(x)\, dx$$

Risk(D(x)=·|x)

Note that we have only options $D(x) \in \{1, \ldots, A\}$ for a given x. The risk of deciding $D(x) = d$ for a given x is as follows:

class posterior for label l

$$Risk(D(x)=d|x) = \sum_{l=1}^{C} \lambda_{dl} \, P(L=l|x)$$

the summation here produces the average loss of deciding d given x, considering the class posterior probabilities for the given x.

loss we would incur by deciding d given sample from class l

Probability of class label being l given x

Since $\lambda_{dl} \geq 0$ and $P(L=l|x) \geq 0$ for all (d,l) pairs, $Risk(D(x)=d|x) \geq 0$ for all $d \in \{1, \ldots, A\}$.

Back to $\underset{\bar{X}}{E}[\text{Risk}] = \int_{-\infty}^{\infty} \overbrace{\text{Risk}(D(x)=\cdot \mid x)}^{\geq 0} \overbrace{p_{\bar{X}}(x)}^{\geq 0} dx$

we notice now that both $\text{Risk}(D(x)=\cdot \mid x) \geq 0$ and $p_X(x) \geq 0$ for all $x \in \mathbb{R}^n$. Also, we notice that $\text{Risk}(D(x)=\cdot \mid x)$ only depends on x (and not on decisions we make for other x'). Consequently, if we make the minimum-risk decision for each individual x, we will end up minimizing $E_{\bar{X}}[\text{Risk}]$ overall.

So, the expected-risk-minimization (ERM) decision rule becomes:

$$D(x) = \underset{d \in \{1,\dots,A\}}{\arg\min} \; \text{Risk}(D(x)=d \mid x)$$

$$= \underset{d \in \{1,\dots,A\}}{\arg\min} \sum_{l=1}^{C} \lambda_{dl} P(L=l \mid x)$$

Bayes Rule

$$= \underset{d \in \{1,\dots,A\}}{\arg\min} \sum_{l=1}^{C} \lambda_{dl} \frac{p(x \mid L=l) P(L=l)}{p(x)}$$

multiply by $p(x)$, which does not change the decision

$$= \underset{d \in \{1,\dots,A\}}{\arg\min} \sum_{l=1}^{C} \lambda_{dl} p(x \mid L=l) P(L=l)$$

Basically, for a given x, we will compute

$$R(D=1|x) = \sum_{l=1}^{C} \lambda_{1l}\, P(L=l|x)$$

$$R(D=2|x) = \sum_{l=1}^{C} \lambda_{2l}\, P(L=l|x)$$

class posteriors

$$\vdots$$

$$R(D=A|x) = \sum_{l=1}^{C} \lambda_{Al}\, P(L=l|x)$$

and choose the decision with the smallest $R(D(x)=\cdot|x)$ value as the best option.

Note that, in vectorized form, these equations can be written as

$$\begin{bmatrix} R(D=1|x) \\ \vdots \\ R(D=A|x) \end{bmatrix} = \Lambda \begin{bmatrix} P(L=1|x) \\ \vdots \\ P(L=C|x) \end{bmatrix}$$

all decision risks across options

loss matrix

all class posteriors

As shown before, the class posteriors can be computed from class conditional pdfs and class priors:

$$P(L=l|x) = \frac{p(x|L=l)P(L=l)}{p(x)}$$

where $p(x) = \sum_{j=1}^{C} p(x|L=j)P(L=j)$

The class-posterior vector is, then

$$\begin{bmatrix} P(L=1|x) \\ \vdots \\ P(L=C|x) \end{bmatrix} = \frac{1}{p(x)} \begin{bmatrix} p(x|L=1)P(L=1) \\ \vdots \\ p(x|L=C)P(L=C) \end{bmatrix}$$

$$= \frac{1}{p(x)} \begin{bmatrix} P(L=1) & & 0 \\ & \ddots & \\ 0 & & P(L=C) \end{bmatrix} \begin{bmatrix} p(x|L=1) \\ \vdots \\ p(x|L=C) \end{bmatrix}$$

Later we will see discriminative models that attempt to approximate class posteriors directly, and generative models that attempt to approximate class conditionals/priors.

Special Case: Minimum Probability of Error Classification Rule

Let $l, d \in \{1, \dots, C\}$ $\Leftrightarrow$ our decisions will come from the same set as class labels

Let $\lambda_{dl} = \begin{cases} 0 & \text{if } d = l \\ 1 & \text{if } d \neq l \end{cases}$ 0 - loss for correct decision, 1 - loss for any incorrect decision

(δ_{dl}: Kronecker delta) $\searrow$ $= 1 - \delta_{dl}$

This is called the 0-1 loss and it indicates our design choice that all incorrect decisions are equally bad and they are worse than all correct decisions which are equally good.

$$R(D = d \mid x) = \sum_{l=1}^{C} \lambda_{dl} P(L = l \mid x)$$

$\lambda_{dl} = 1 - \delta_{dl}$ $\downarrow$

$$= \sum_{\substack{l=1 \\ l \neq d}}^{C} P(L = l \mid x)$$

$$= 1 - P(L = d \mid x)$$

After this substitution, the ERM decision rule simplifies to the Maximum A Posteriori (MAP) decision rule:

$$D(x) = \underset{d \in \{1, \dots, C\}}{\operatorname{argmin}} \; R(D = d \mid x)$$

$$= \underset{d \in \{1, \dots, C\}}{\operatorname{argmin}} \; 1 - P(L = d \mid x)$$

$$= \underset{d \in \{1, \dots, C\}}{\operatorname{argmax}} \; P(L = d \mid x)$$

For a given x, decide on the class label with the largest posterior probability.

MAP classification rule (i.e. ERM with 0-1 loss) achieves minimum probability of error overall.

Special Case of MAP Classifier: ML Classifier

If the class priors are equal, then MAP classifier simplifies to ML classifier

ML: maximum likelihood.

$$D_{MAP}(x) = \underset{d \in \{1, \dots, C\}}{\arg\max}\ P(L=d \mid x)$$

$$= \underset{d}{\arg\max}\ \frac{p(x \mid L=d)\, P(L=d)}{p(x)}$$

multiply with $p(x)$

$$= \underset{d}{\arg\max}\ p(x \mid L=d)\, P(L=d)$$

$P(L=d) = \frac{1}{C}$ for all d multiply with C

$$= \underset{d \in \{1, \dots, C\}}{\arg\max}\ p(x \mid L=d)$$

select the class label that has a class cond. pdf which makes x most likely.

If you want to minimize probability of error, use 0-1 loss ($\Rightarrow$ MAP). In that case if class priors are equal, you can simply use this ML classification rule.

Special Case: ERM with 2 classes.

Let $l, d \in \{0, 1\}$. Then

$$D_{ERM}(x) = \underset{d \in \{0,1\}}{\operatorname{argmin}} \; R(D = d \mid x)$$

$$= \underset{d \in \{0,1\}}{\operatorname{argmin}} \; \lambda_{d0} P(L=0 \mid x) + \lambda_{d1} P(L=1 \mid x)$$

i.e. we will decide as follows

$$\lambda_{00} P(L=0 \mid x) + \lambda_{01} P(L=1 \mid x) \underset{D(x)=0}{\overset{D(x)=1}{\gtrless}} \lambda_{10} P(L=0 \mid x) + \lambda_{11} P(L=1 \mid x)$$

$\downarrow$ Let $\lambda_{01} - \lambda_{11} > 0$ and $\lambda_{10} - \lambda_{00} > 0$

(i.e. incorrect decisions are worse than correct decisions)

Then

$$\frac{P(L=1 \mid x)}{P(L=0 \mid x)} \underset{D(x)=0}{\overset{D(x)=1}{\gtrless}} \frac{(\lambda_{10} - \lambda_{00})}{(\lambda_{01} - \lambda_{11})}$$

Decide based on the ratio of class posteriors

$\hookrightarrow$ (if 0-1 loss threshold $= 1$)

$\downarrow$ with Bayes rule $|||$

$$\frac{p(x \mid L=1)}{p(x \mid L=0)} \underset{D(x)=0}{\overset{D(x)=1}{\gtrless}} \frac{(\lambda_{10} - \lambda_{00}) P(L=0 \cancel{\mid x})}{(\lambda_{01} - \lambda_{11}) P(L=1 \cancel{\mid x})}$$

Decide based on the ratio of likelihoods under class cond pdfs.