The Classification Problem and Statistical Framework

Andrew Nobel

September, 2023

Classification

Data: Labeled pairs $(x_1, y_1), \ldots, (x_n, y_n)$ with

- $x_i \in \mathcal{X}$ space of *predictors* (often $\mathcal{X} \subseteq \mathbb{R}^d$)
- $y_i \in \{0, 1\}$ response or *class label*

Goal: Given an *unlabeled* predictor $x \in \mathcal{X}$, assign it to class 0 or 1

- Label may be unavailable or expensive to obtain

Idea: Use labeled examples to classify unlabeled ones

Example: Spam Recognition

Predictor: $x =$ vector of features extracted from text of email, e.g.,

- presence of keywords (“cheap”, “cash”, “medicine”)
- presence of key phrases (“Dear Sir/Madam”)
- use of words in all-caps (“VIAGRA”)
- point of origin of email

Response: $y = 1$ if email is spam, $y = 0$ otherwise

Task: Given sample $(x_1, y_1), \ldots, (x_n, y_n)$ of labeled emails, construct a prediction rule to classify future email messages as spam or not-spam

Measuring Errors in Prediction

Definition: A *classification rule* is a map $\phi : \mathcal{X} \to \{0, 1\}$. Regard $\phi(x)$ as a prediction of the class label associated with x

Zero-One loss: Performance of ϕ on pair (x, y) given by

$$\ell(\phi(x), y) \;=\; \mathbb{I}(\phi(x) \neq y) \;=\; \begin{cases} 1 & \text{if } \phi(x) \neq y \\ 0 & \text{if } \phi(x) = y \end{cases}$$

Summary table: For (x, y) pair four possible outcomes

	$\phi(x) = 1$	$\phi(x) = 0$
$y = 1$	correct	error
$y = 0$	error	correct

Receiver Operating Characteristic (ROC) Curves

Idea: Diagram to assess performance of a family of classification rules, usually parametrized by a fixed threshold.

Setting: Binary classification with two outcomes

- $1 =$ "positive"
- $0 =$ "negative"

Confusion Matrix: For rule $\phi : \mathcal{X} \to \{0, 1\}$ and data $(x_1, y_1), \ldots, (x_n, y_n)$ we can summarize outcome of predictions as follows

	$\phi = 1$	$\phi = 0$
$y = 1$	true positives	false negatives
$y = 0$	false positives	true negatives

ROC Curves, cont.

Defn: Given rule $\phi : \mathcal{X} \to \{0, 1\}$ and data $(x_1, y_1), \ldots, (x_n, y_n)$

- True positive rate (Sensitivity)

$$\text{TPR}(\phi) \;=\; \frac{\sum_i \phi(x_i) y_i}{\sum_i y_i} \;=\; \frac{\text{\# true positive predictions}}{\text{total \# positives}}$$

- True negative rate (Specificity)

$$\text{TNR}(\phi) \;=\; \frac{\sum_i (1 - \phi(x_i))(1 - y_i)}{\sum_i (1 - y_i)} \;=\; \frac{\text{\# true negative predictions}}{\text{total \# negatives}}$$

- False positive/alarm rate

$$\text{FPR}(\phi) = 1 - \text{TNR}(\phi) \;=\; \frac{\text{\# false positive predictions}}{\text{total \# negatives}}$$

ROC Curve

Given: Ordered family $\mathcal{F} = \{\phi_t : t \in T\}$ of classification rules, e.g.,

$$\phi_t(x) = \mathbb{I}(x \geq t) \;\text{ or }\; \phi_t(x) = \mathbb{I}(\langle x, v \rangle \geq t)$$

Note: decreasing t increases both false and true positive rates.

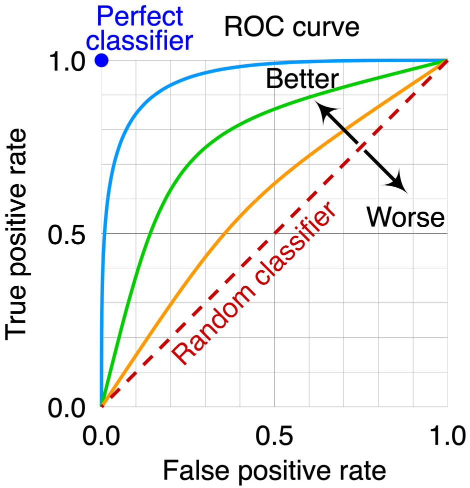
Definition: ROC curve of the family $\mathcal{F}$ is a plot of

$$(\text{FPR}(\phi_t), \text{TPR}(\phi_t)) \in [0,1]^2 \;\text{ for }\; t \in T$$

Ideally TPR(ϕ) is close to one when FPR(ϕ) is close to zero

AUC: Quality of family $\mathcal{F}$ assessed by area under the ROC curve

ROC Illustration (cmglee, from Wikipedia)



Decision Regions and Decision Boundary

Note: Every rule $\phi : \mathcal{X} \to \{0, 1\}$ partitions $\mathcal{X}$ into two sets

$$\mathcal{X}_0(\phi) = \{x \in \mathcal{X} : \phi(x) = 0\}$$

$$\mathcal{X}_1(\phi) = \{x \in \mathcal{X} : \phi(x) = 1\}$$

Terminology

- Sets $\mathcal{X}_0(\phi), \mathcal{X}_1(\phi)$ called *decision regions* of ϕ
- Interface between $\mathcal{X}_0(\phi)$ and $\mathcal{X}_1(\phi)$ called *decision boundary* of ϕ

Classification Problem Revisited

Picture

- Write sample $(x_1, y_1), \ldots, (x_n, y_n)$ as points $x_i \in \mathcal{X}$ with labels y_i
- Look for decision regions that (mostly) separate zeros and ones

Two Related Issues

- Tradeoff between complexity and separation
- Will selected rule perform well on future, unlabeled, samples?

The Stochastic Setting

Stochastic Setting

Assumptions

- Observations $D_n = (X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \{0, 1\}$ random
- (X_i, Y_i) drawn independently from distribution P on $\mathcal{X} \times \{0, 1\}$
- Future observation (X, Y) drawn independently from same distribution P

Key Stochastic Quantities

1. Prior probabilities of $Y = 0$ and $Y = 1$
2. Conditional probability of $Y = 1$ given $X = x$
3. Conditional distribution of X given $Y = 0$ and $Y = 1$

Prior Probabilities

Given: Joint pair $(X, Y) \in \mathcal{X} \times \{0, 1\}$

Define: Prior probabilities $\pi_0 = \mathbb{P}(Y = 0)$ and $\pi_1 = \mathbb{P}(Y = 1)$

Notes

- Probability of seeing class $Y = 0$ or $Y = 1$ *prior* to observing X
- π_0, π_1 represent relative abundance of class 0 and 1
- Note that $\pi_0 + \pi_1 = 1$
- Cases in which $\pi_1 >> \pi_0$ or vice versa can be difficult

Unconditional and Conditional Densities of X

Given: Joint pair $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$

Define: Unconditional and conditional densities of X

- $f(x) =$ *unconditional density* of X

$$\mathbb{P}(X \in A) \; = \; \int_A f(x) \, dx \qquad A \subseteq \mathcal{X}$$

- $f(x|0), f(x|1) =$ *class-conditional densities* of X

$$\mathbb{P}(X \in A \,|\, Y = y) \; = \; \int_A f(x|y) \, dx \qquad A \subseteq \mathcal{X}$$

Note: Densities $f(\cdot|0)$ and $f(\cdot|1)$ tell us about separability of 0s and 1s

Conditional Distribution of Y Given X

Given: Joint pair $(X, Y) \in \mathcal{X} \times \{0, 1\}$

Define: Conditional probability $\eta(x) \;=\; \mathbb{P}(Y = 1 \,|\, X = x)$

- Posterior probability that $Y = 1$ given that $X = x$
- Note that $\mathbb{P}(Y = 0 \,|\, X = x) = 1 - \eta(x)$.

Regimes:

- $\eta(x) \approx 1 \;\Rightarrow Y$ is likely to be 1 given $X = x$
- $\eta(x) \approx 0 \;\Rightarrow Y$ is likely to be 0 given $X = x$
- $\eta(x) \approx 1/2 \;\Rightarrow$ value of Y uncertain given $X = x$

Relations Among Distributions

1. By the law of total probability we have

$$f(x) \ = \ \pi_0 \, f(x|0) \ + \ \pi_1 \, f(x|1)$$

Moreover, as f_0 and f_1 are densities $\int f(x|0)dx = \int f(x|1)dx = 1$

2. By Bayes theorem we know

$$\eta(x) \ = \ \frac{\pi_1 \, f(x|1)}{f(x)} \ = \ \frac{\pi_1 \, f(x|1)}{\pi_0 \, f(x|0) \ + \ \pi_1 \, f(x|1)}$$

Risk of a Prediction Rule

Recall: Performance of rule ϕ on single pair (x, y) given by zero-one loss

$$\ell(\phi(x), y) \;=\; \mathbb{I}(\phi(x) \neq y) \;=\; \begin{cases} 1 & \text{if } \phi(x) \neq y \\ 0 & \text{if } \phi(x) = y \end{cases}$$

Definition: The *risk* of a fixed classification rule ϕ on a random pair (X, Y) is its *expected loss*

$$R(\phi) \;=\; \mathbb{E}[\mathbb{I}(\phi(X) \neq Y)] \;=\; \mathbb{P}(\phi(X) \neq Y)$$

which is just the probability that ϕ misclassifies X

Optimality and the Bayes Rule

Bayes Rule and Bayes Risk

Definition: The *Bayes classification rule* ϕ^* for the pair (X, Y) is

$$\phi^*(x) = \underset{k=0,1}{\operatorname{argmax}}\ \mathbb{P}(Y = k \,|\, X = x)$$

- $\phi^*(x)$ is the most likely value of Y given $X = x$
- $\phi^*(x)$ depends on distribution of (X, Y), usually unknown

Definition: The *Bayes risk* R^* for (X, Y) is the risk of the Bayes rule

$$R^* \;=\; R(\phi^*) \;=\; \mathbb{P}(\phi^*(X) \neq Y)$$

Optimality of the Bayes Rule

Note: For binary Y the Bayes rule has the equivalent forms

$$\phi^*(x) \;=\; \mathbb{I}(\eta(x) \geq 1/2) \;=\; \operatorname*{argmax}_{y=0,1} \pi_y f(x|y)$$

Theorem: The Bayes rule ϕ^* for (X, Y) is optimal: for every classification rule $\phi : \mathcal{X} \to \{0, 1\}$ we have $R^* \leq R(\phi)$.

Fact: The Bayes risk R^* can be written in the form

$$R^* = \mathbb{E} \min\{\eta(X), 1 - \eta(X)\}$$

Understanding the Bayes Risk

Fact: Let $(X, Y) \in \mathcal{X} \times \{0, 1\}$ be a jointly distributed pair

1. Bayes risk $R^* \in [0, 1/2]$
2. $R^* = 0$ iff $\eta(x) \in \{0, 1\}$ iff Y is a function of X
3. $R^* = 1/2$ iff $\eta(x) \equiv 1/2$ which implies that $Y \perp\!\!\!\perp X$
4. If $Y \perp\!\!\!\perp X$ then $\phi^*(x)$ is constant (1 if $\pi_1 \geq \pi_0$ and 0 if $\pi_0 < \pi_1$)

Fixed vs. Data Dependent Prediction Rules

Observations $D_n = (X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \{0, 1\}$ iid $\sim (X, Y)$

Fixed rule $\phi : \mathcal{X} \to \{0, 1\}$

- $\phi(x)$ predicts class label of x without regard to D_n
- Risk $R(\phi) = \mathbb{P}(\phi(X) \neq Y)$ is a constant

Classification procedure $\phi_n : \mathcal{X} \times (\mathcal{X} \times \{0, 1\})^n \to \{0, 1\}$

- $\hat{\phi}_n(x) = \phi_n(x : D_n)$ predicts class label of x *based on* D_n
- Risk $R(\hat{\phi}_n) = \mathbb{P}(\hat{\phi}_n(X) \neq Y \mid D_n)$ is a random variable

Classification Procedures Based on Distributional Assumptions

Linear and Quadratic Discriminant Analysis

Idea: Assume class conditional density $f(x|y) = \mathcal{N}_d(\mu_y, \Sigma_y)$ for $y = 0, 1$

Fitting: Given observations D_n

1. Estimate mean $\hat{\mu}_y$ and variance $\hat{\Sigma}_y$. Let $\hat{f}(x|y) = \mathcal{N}_d(\hat{\mu}_y, \hat{\Sigma}_y)$
2. Estimate priors $\hat{\pi}_y$
3. Define $\hat{\phi}(x) = \mathrm{argmax}_{y=0,1}\, \hat{\pi}_y \hat{f}(x|y)$

1. LDA: Assume $\Sigma_0 = \Sigma_1$. In this case $\hat{\phi}$ has linear decision boundary

2. QDA: Allow $\Sigma_0 \neq \Sigma_1$. In this case $\hat{\phi}$ has quadratic decision boundary

Logistic Regression Model

Model: For some coefficient vector $\beta \in \mathbb{R}^{d+1}$

$$\log \frac{\eta(x)}{1-\eta(x)} \;=\; \langle \beta, x \rangle \quad \text{equivalently} \quad \eta(x : \beta) \;=\; \frac{e^{\langle \beta, x \rangle}}{1 + e^{\langle \beta, x \rangle}}$$

Fitting: Given observations D_n find the coefficient vector $\hat{\beta}$ maximizing the *conditional log-likelihood* (using gradient descent)

$$\ell(\beta) \;=\; \log \prod_{i=1}^{n} \eta(x_i : \beta)^{y_i} \, (1 - \eta(x_i : \beta))^{1-y_i}$$

Define rule $\hat{\phi}(x) = \mathbb{I}(\eta(x : \hat{\beta}) \geq 1/2)$

Naive Bayes

Setting: Covariate $X = (X_1, \ldots, X_d)^t$ has d components. Assume the components are conditionally independent given Y: for $y = 0, 1$

$$f(x_1, \ldots, x_d \,|\, y) \;=\; f_1(x_1 \,|\, y) \cdots f_d(x_d \,|\, y)$$

Approach: Given observations D_n

- Form estimates $\hat{f}_j(x_j \,|\, y)$ of conditional marginals
- Estimate class conditional $\hat{f}(x \,|\, y) = \prod_{j=1}^d \hat{f}_j(x_j \,|\, y)$
- Combine with estimates $\hat{\pi}_0, \hat{\pi}_1$ of priors to obtain the rule

$$\hat{\phi}(x) \;=\; \underset{j=0,1}{\operatorname{argmax}} \, \hat{\pi}_y \, \hat{f}(x \,|\, y)$$

More General Classification Procedures

Histogram Rules

- Observations $D_n = (X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \{0, 1\}$
- Partition $\pi = \{A_1, \ldots, A_K\}$ of $\mathcal{X}$ into disjoint sets called cells
- Let $\pi(x) =$ cell A_k of π containing x

Definition: The histogram classification rule for π is given by

$$\phi_n^\pi(x : D_n) \; = \; \hat{\phi}_n^\pi(x) \; = \; \text{maj-vote}\{Y_i : X_i \in \pi(x)\}$$

- Classifies x using "local" data in the same cell as x
- No assumptions about the distribution of (X, Y)
- Decision regions of rule determined by cells of π

Nearest Neighbor Rules

Setting: Observations $D_n = (X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \{0, 1\}$.

For $x \in \mathbb{R}^d$ let $X_{(1)}(x), \ldots, X_{(n)}(x)$ be reordering of $X_1, \ldots, X_n$ s.t.

$$||x - X_{(1)}(x)|| \leq ||x - X_{(2)}(x)|| \leq \cdots \leq ||x - X_{(n)}(x)||$$

Let $Y_{(j)}(x) =$ class label of $X_{(j)}(x) =$ the jth nearest neighbor of x.

Definition: For $k \geq 1$ odd the *k-nearest neighbor rule* is given by

$$\phi_n^{\text{k-NN}}(x) \ = \ \text{majority-vote}\{Y_{(1)}(x), \ldots, Y_{(k)}(x)\}$$

Asymptotic Performance of 1-NN Rule

Note: The 1-NN rule assigns to $x \in \mathbb{R}^d$ the label of the nearest X_i

Theorem (Cover and Hart) As the number of samples n tends to infinity,

$$\mathbb{E}R(\hat{\phi}_n^{\text{1-NN}}) \;\to\; 2\mathbb{E}[\eta(X)(1-\eta(X))] \;\leq\; 2R^*$$

Upshot: asymptotic probability of error of 1-NN rule is *at most* twice the Bayes risk (best performance of any classification rule)!

Other Methods

- Classification trees
- Bagging
- Boosting
- Support vector machines