PREDICTION PROBLEMS

CS6140

Predrag Radivojac
KHOURY COLLEGE OF COMPUTER SCIENCES
NORTHEASTERN UNIVERSITY

Fall 2024

Types of Problems in Machine Learning

Some buzzwords frequently mentioned:

1. Supervised learning
2. Unsupervised learning
3. Semi-supervised learning
4. Completion under mising features
5. Learning to rank
6. Statistical relational learning
7. Active learning
8. Structured prediction
9. Reinforcement learning
10. Online learning

And more.

These are not mutually exclusive.

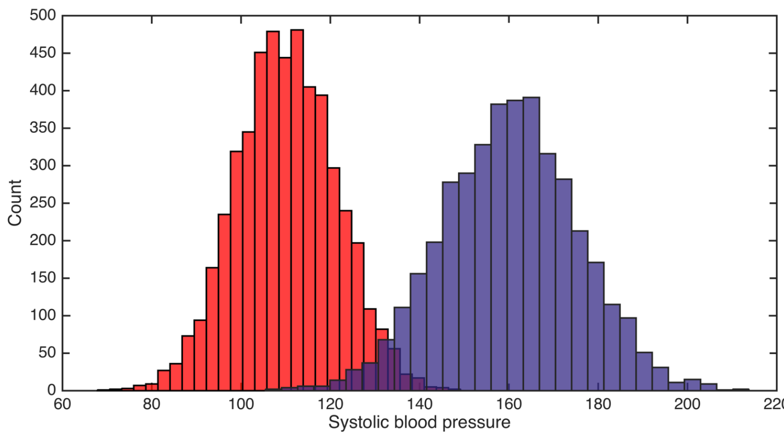
Supervised Learning (Classification)

Given:

$\mathcal{D}_{\text{red}}$: sample from people w/o heart disease

$\mathcal{D}_{\text{blue}}$: sample from people w/ heart disease

Goal: predict heart disease



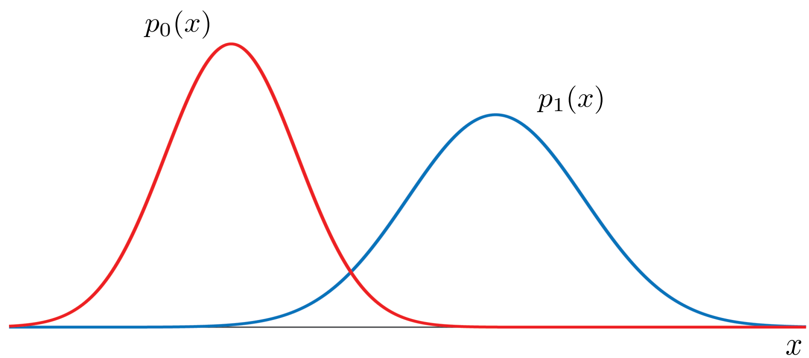
$x \in \mathbb{R}$

$y \in \{\text{disease}, \text{no disease}\}$

Supervised Learning

$\mathcal{D}_0$ = sample from $p_0(x)$

$\mathcal{D}_1$ = sample from $p_1(x)$



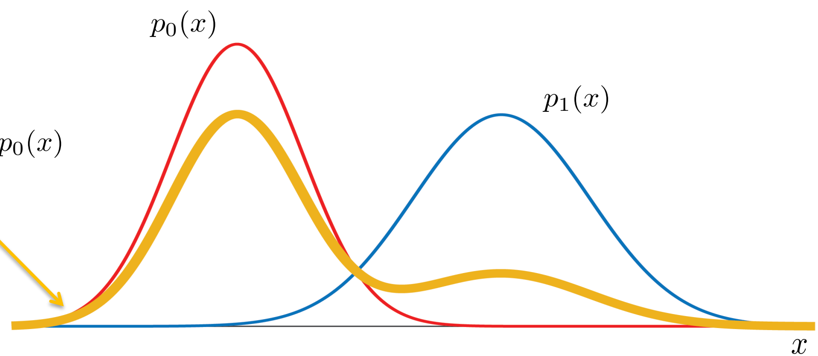
Semi-supervised Learning

$\mathcal{D}_0$ = sample from $p_0(x)$

$\mathcal{D}_1$ = sample from $p_1(x)$

$\mathcal{D}$ = sample from $p(x)$

$$p(x) = \alpha \cdot p_1(x) + (1 - \alpha) \cdot p_0(x)$$



$\alpha \in (0, 1)$, here $\alpha = 0.25$

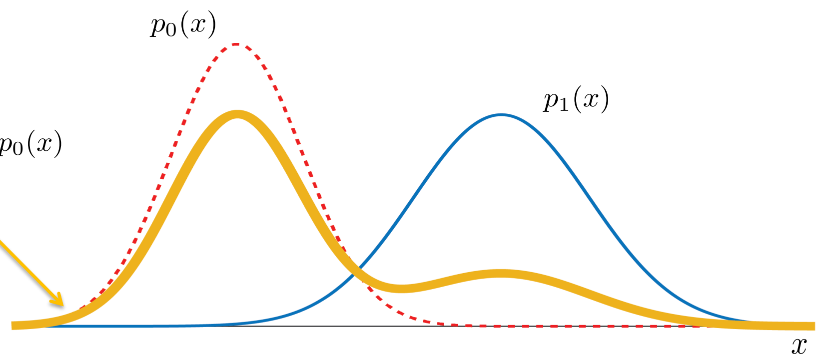
POSITIVE-UNLABELED LEARNING

$\mathcal{D}_0 = \text{sample from } p_0(x)$

$\mathcal{D}_1 = \text{sample from } p_1(x)$

$\mathcal{D} = \text{sample from } p(x)$

$p(x) = \alpha \cdot p_1(x) + (1 - \alpha) \cdot p_0(x)$



$\alpha \in (0, 1)$, here $\alpha = 0.25$

Noisy Positive-Unlabeled Learning

$\mathcal{D}_0$ = sample from $p_0(x)$

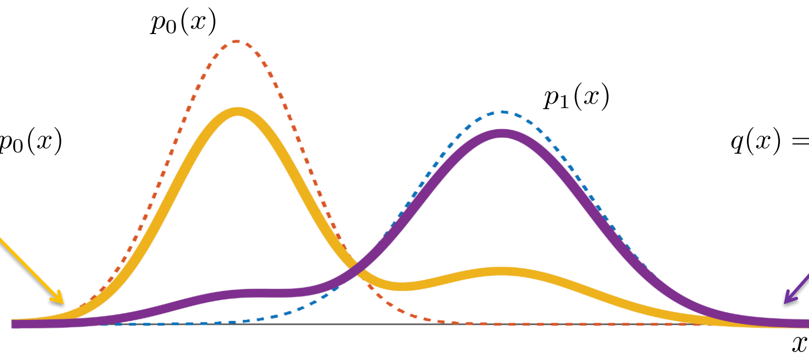
$\mathcal{D}_1$ = sample from $p_1(x)$

$\mathcal{D}$ = sample from $p(x)$

$\mathcal{L}$ = sample from $q(x)$

$p(x) = \alpha \cdot p_1(x) + (1 - \alpha) \cdot p_0(x)$

$q(x) = \beta \cdot p_1(x) + (1 - \beta) \cdot p_0(x)$



$\alpha \in (0, 1)$, here $\alpha = 0.25$

$\beta \in (\alpha, 1]$, here $\beta = 0.90$

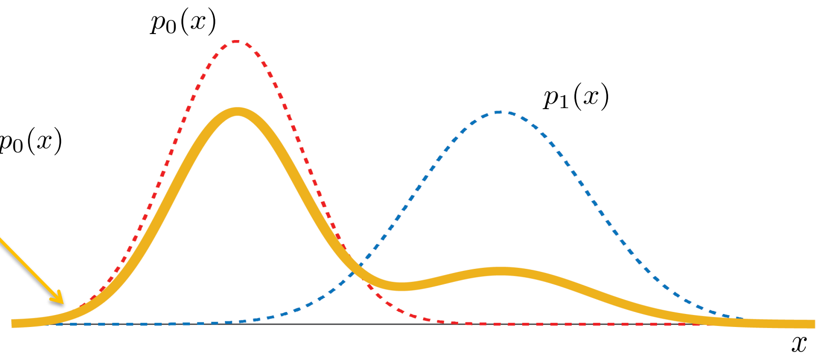
UNSUPERVISED LEARNING

$\mathcal{D}_0$ = sample from $p_0(x)$

$\mathcal{D}_1$ = sample from $p_1(x)$

$\mathcal{D}$ = sample from $p(x)$

$$p(x) = \alpha \cdot p_1(x) + (1-\alpha) \cdot p_0(x)$$



$\alpha \in (0,1)$, here $\alpha = 0.25$

Supervised Learning

Given: $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$

$\boldsymbol{x}_i \in \mathcal{X}$ is the i-th input example (data point, instance, object, pattern)
$y_i \in \mathcal{Y}$ is the i-th target value
$\mathcal{X}$ = input space, often $\mathbb{R}^d$
$\mathcal{Y}$ = output space

Objective: learn a good mapping $f : \mathcal{X} \to \mathcal{Y}$

- often learn an intermediate mapping $s : \mathcal{X} \to \mathbb{R}$ (classification)

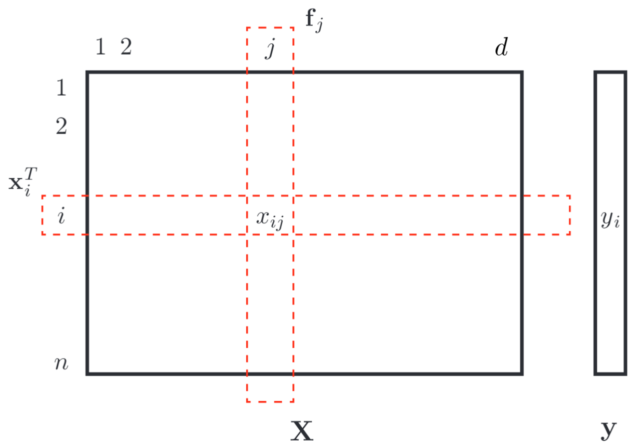
When $\mathcal{X} = \mathbb{R}^d$, we have $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$.

Each dimension of $\boldsymbol{x}$ is called a feature or attribute.

Each x_j is called a feature or attribute value.

Vector Space Representation

We often have the following setup:



$\mathbf{X} = n \times d$ data (design) matrix $\quad\quad$ $\mathbf{y} = n \times 1$ target (response) vector

CLASSIFICATION

$\mathcal{Y}$ is discrete

Consider a problem of predicting a disease state of an individual.

$$\mathcal{Y} = \{-1, +1\}$$

	wt [kg]	ht [m]	T [°C]	sbp [mmHg]	dbp [mmHg]	y
$\boldsymbol{x}_1$	91	1.85	36.6	121	75	-1
$\boldsymbol{x}_2$	75	1.80	37.4	128	85	$+1$
$\boldsymbol{x}_3$	54	1.56	36.6	110	62	-1

$\boldsymbol{X}$ = descriptors of each individual
Y = the disease state for each individual

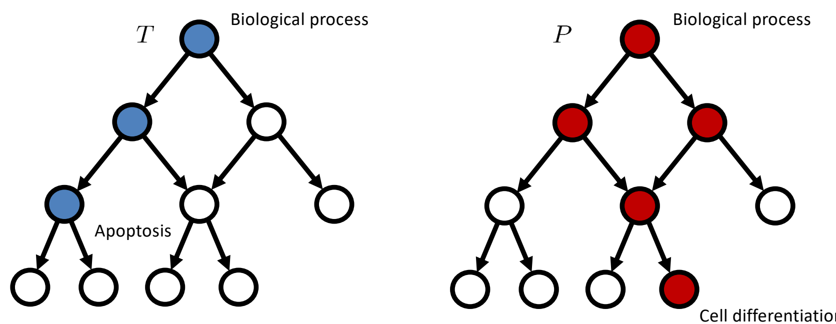
Types of Classification

Binary: $\mathcal{Y} = \{\text{spam}, \ \text{not spam}\}$

Multi-class: $\mathcal{Y} = \{\mathrm{A}, \mathrm{B}, \mathrm{AB}, \mathrm{O}\}$

Multi-label: consider categories $\{\text{sports}, \ \text{medicine}, \ \text{travel}, \ \text{politics}\}$

Structured-output:



Regression

$\mathcal{Y}$ is continuous

Consider a problem of predicting the price of a house.

$$\mathcal{Y} = [0, \infty)$$

	size [sqft]	age [yr]	dist [mi]	inc [$]	dens [ppl/mi^2]	y
$\boldsymbol{x}_1$	1250	5	2.85	56,650	12.5	2.35
$\boldsymbol{x}_2$	3200	9	8.21	245,800	3.1	3.95
$\boldsymbol{x}_3$	825	12	0.34	61,050	112.5	5.10

$\boldsymbol{X}$ = descriptors of each house
Y = the price a house is sold at in $100k

Optimal Classification

Suppose $p(\boldsymbol{x}, y)$ is known, $c : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ is some cost function (matrix).

$$\begin{aligned} \mathbb{E}[C] &= \int_{\mathcal{X}} \sum_{y} c(\hat{y}, y) p(\boldsymbol{x}, y) d\boldsymbol{x} \\ &= \int_{\mathcal{X}} p(\boldsymbol{x}) \sum_{y} c(\hat{y}, y) p(y|\boldsymbol{x}) d\boldsymbol{x} \end{aligned}$$

Expected cost

Note: $\hat{y} = f(\boldsymbol{x})$

A classifier that minimizes this is

$$f_{\mathrm{BR}}(\boldsymbol{x}) = \operatorname*{arg\,min}_{\hat{y} \in \mathcal{Y}} \left\{ \sum_{y} c(\hat{y}, y) p(y|\boldsymbol{x}) \right\}$$

Bayes risk classifier

Optimal Classification

Minimizing the probability of a classifier's error $P(f(\boldsymbol{x}) \neq y)$

$$c(\hat{y}, y) = \begin{cases} 0 & \text{when } y = \hat{y} \\ 1 & \text{when } y \neq \hat{y} \end{cases}$$

Cost to minimize error

A classifier that minimizes the probability of error:

$$f_{\text{MAP}}(\boldsymbol{x}) = \underset{y \in \mathcal{Y}}{\arg\max} \{p(y|\boldsymbol{x})\}.$$

MAP classifier

$\boxed{\text{Minimizing error is the same as accurately learning posterior distributions } p(y|\boldsymbol{x})}$

MODELING

Well, it comes down to learning $p(y|\boldsymbol{x})$. Assume discrete $\mathcal{Y}$.

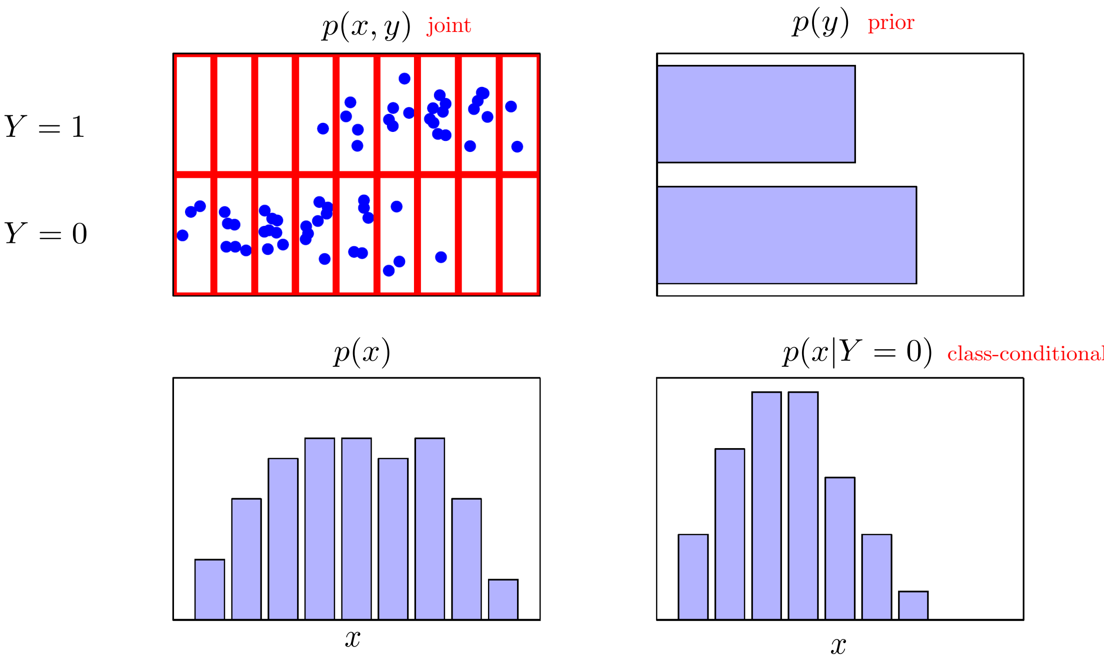
$$\begin{aligned} p(y|\boldsymbol{x}) &= \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x})} \\ &= \frac{p(\boldsymbol{x}|y)p(y)}{\sum_y p(\boldsymbol{x}, y)} \\ &= \frac{p(\boldsymbol{x}|y)p(y)}{\sum_y p(\boldsymbol{x}|y)p(y)} \end{aligned}$$

Learn $p(y|\boldsymbol{x}) \rightarrow$ discriminative model (often assumes data comes from $p(\boldsymbol{x})$).

Learn $p(\boldsymbol{x}|y)$ and $p(y) \rightarrow$ generative model.

One does not need to explicitly learn in either of these ways.

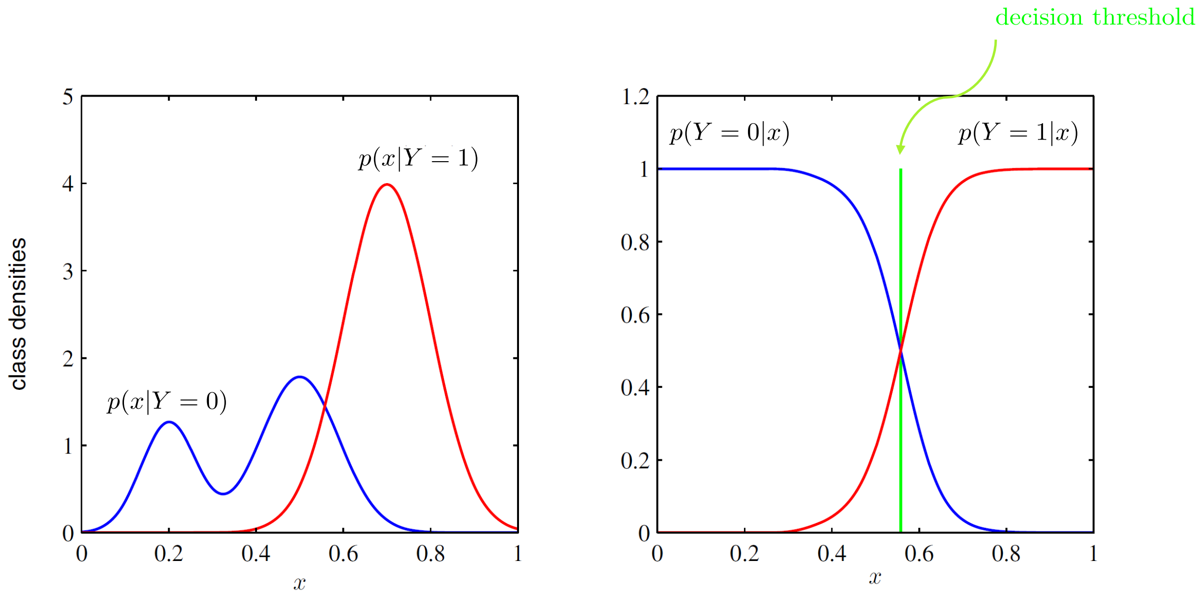
Modeling



$p(Y = 1|x)$ posterior

Picture modified from Bishop's textbook.

Decision Making



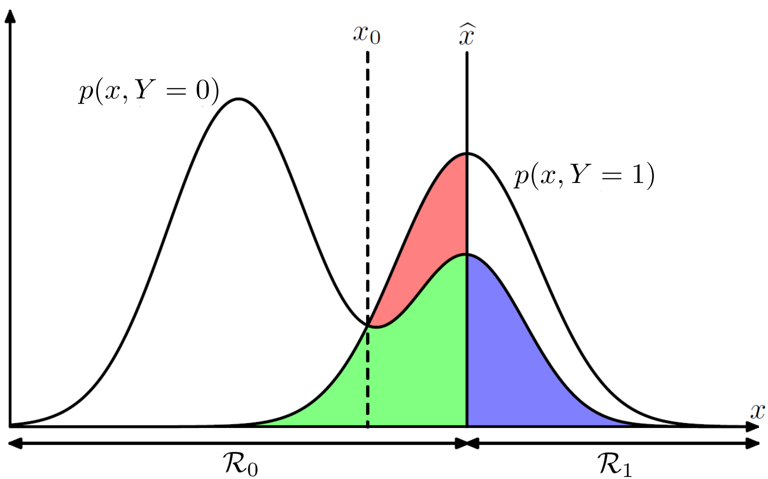
Picture modified from Bishop's textbook.

Decision Making

$$\mathbb{E}[C] = \int_{\mathcal{X}} \sum_y c(\hat{y}, y) p(\boldsymbol{x}, y) d\boldsymbol{x}$$
$$= \int_{\mathcal{R}_0} p(\boldsymbol{x}, Y=1) d\boldsymbol{x} + \int_{\mathcal{R}_1} p(\boldsymbol{x}, Y=0) d\boldsymbol{x}$$

Costs are 0 for correct and 1 for incorrect classification.

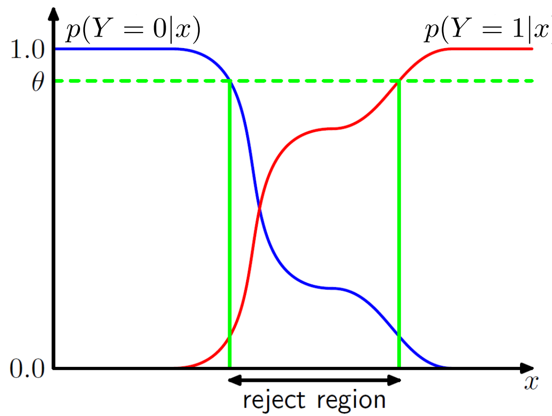
$\mathcal{R}_0$: $\hat{y} = 0$
$\mathcal{R}_1$: $\hat{y} = 1$



$\hat{x}$: our decision threshold
x_0: optimal decision threshold

Picture modified from Bishop's textbook.

Classification with Rejection



Picture modified from Bishop's textbook.

Optimal Regression

Suppose $p(\boldsymbol{x}, y)$ is known, $c : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ is some cost function.

$$\mathbb{E}[C] = \int_{\mathcal{X}} \int_{\mathcal{Y}} c(f(\boldsymbol{x}), y) p(\boldsymbol{x}, y) dy d\boldsymbol{x}$$

Expected cost

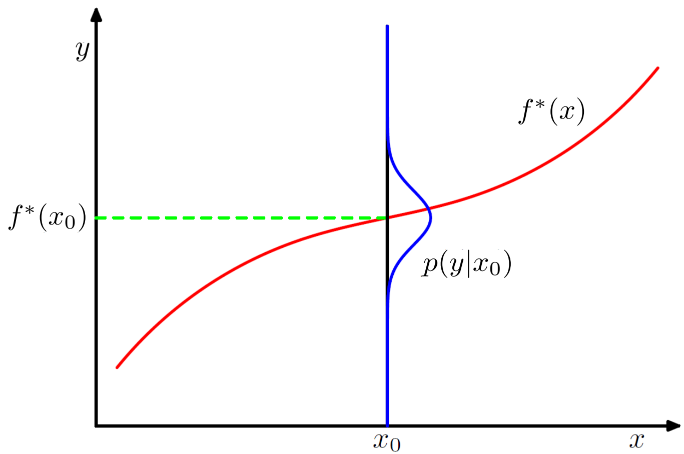
Take $c(f(\boldsymbol{x}), y) = (f(\boldsymbol{x}) - y)^2$. We can now derive

$$\begin{aligned} f^*(\boldsymbol{x}) &= \int_{\mathcal{Y}} y p(y|\boldsymbol{x}) dy \\ &= \mathbb{E}[Y|\boldsymbol{x}] \end{aligned}$$

Optimal regression

Optimal Regression

$\mathbb{E}[Y|x] = \text{optimal regression model}$



Picture modified from Bishop's textbook.

Optimal Regression For L_2 Loss

When $c(f(\boldsymbol{x}), y) = (f(\boldsymbol{x}) - y)^2$, the error decomposes to

$$
\begin{aligned}
\mathbb{E}[C] &= \int_{\mathcal{X}} \int_{\mathcal{Y}} (f(\boldsymbol{x}) - y)^2 p(\boldsymbol{x}, y) dy d\boldsymbol{x} \\
&= \int_{\mathcal{X}} (f(\boldsymbol{x}) - \mathbb{E}[Y|\boldsymbol{x}])^2 p(\boldsymbol{x}) d\boldsymbol{x} + \int_{\mathcal{X}} \int_{\mathcal{Y}} (\mathbb{E}[Y|\boldsymbol{x}] - y)^2 p(\boldsymbol{x}, y) dy d\boldsymbol{x}
\end{aligned}
$$

Reducible error

Irreducible error

See lecture notes for complete derivation.

Bias-Variance Tradeoff

Consider the reducible error (RE) term

$$\int_{\mathcal{X}} (f(\boldsymbol{x}) - \mathbb{E}[Y|\boldsymbol{x}])^2 p(\boldsymbol{x}) d\boldsymbol{x}$$

Consider further

1. the predictor depends on $\mathcal{D}$; i.e., $f(\boldsymbol{x}) \to f(\boldsymbol{x}|\mathcal{D})$
2. $\mathcal{D}$ is a realization of random variable D; i.e., $f(\boldsymbol{x}|D)$ is too
3. we can look at the expectation of $f(\boldsymbol{x}|\mathcal{D})$; i.e., $\mathbb{E}[f(\boldsymbol{x}|D)]$

Bias-Variance Tradeoff

The expected Reducible Error (RE), wrt random variable D

$$\begin{aligned}\text{Expected RE} &= \mathbb{E}\left[\int_{\mathcal{X}} (f(\boldsymbol{x}|D) - \mathbb{E}[Y|\boldsymbol{x}])^2 p(\boldsymbol{x}) d\boldsymbol{x}\right] \\ &= \underbrace{\int_{\mathcal{X}} \left(\mathbb{E}[f(\boldsymbol{x}|D)] - \mathbb{E}[Y|\boldsymbol{x}]\right)^2 p(\boldsymbol{x}) d\boldsymbol{x}}_{\text{bias}^2} + \underbrace{\int_{\mathcal{X}} \mathbb{E}\left[\left(f(\boldsymbol{x}|D) - \mathbb{E}[f(\boldsymbol{x}|D)]\right)^2\right] p(\boldsymbol{x}) d\boldsymbol{x}}_{\text{variance}}\end{aligned}$$

Bias: how much the expected output deviates from the optimal

Variance: how much the output deviates from its expected value

Derivation in Bishop's textbook (Chapter 3).

Naive Bayes Model

Given: a set of observations $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, $x_i \in \mathcal{X}, y_i \in \mathcal{Y}$

Objective: learn the posterior $p(y|x, \mathcal{D})$

Naive Bayes Model:

$$\begin{aligned} p(y|\boldsymbol{x}) &= \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x})} \\ &= \frac{p(\boldsymbol{x}|y)p(y)}{\sum_y p(\boldsymbol{x}, y)} \\ &= \frac{p(\boldsymbol{x}|y)p(y)}{\sum_y p(\boldsymbol{x}|y)p(y)} \end{aligned}$$

Assume $\mathcal{X} = \mathbb{R}^d$.

Assume discrete $\mathcal{Y}$.

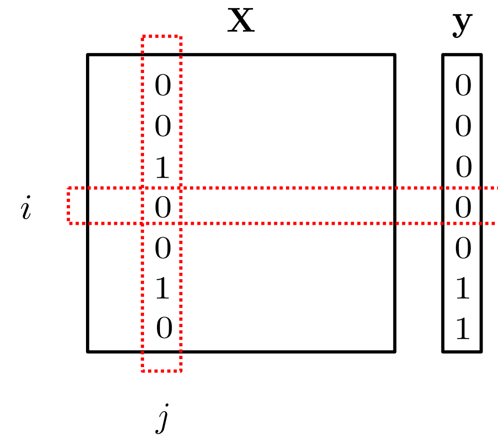
$$p(x_1, x_2, ..., x_d|y) = \prod_{j=1}^{d} p(x_j|y)$$

$\leftarrow$ naive Bayes assumption

Naive Bayes Classification

Assume: discrete $\mathcal{X}_j$, discrete $\mathcal{Y}$

generalized Bernoulli distribution



$$\forall i \quad x_{i,j} = l \in \mathcal{X}_j$$

$$y_i = k \in \mathcal{Y}$$

$$P(X_j = l | Y = k) = \alpha_{j,l,k}$$

$$P(Y = k) = \alpha_k$$

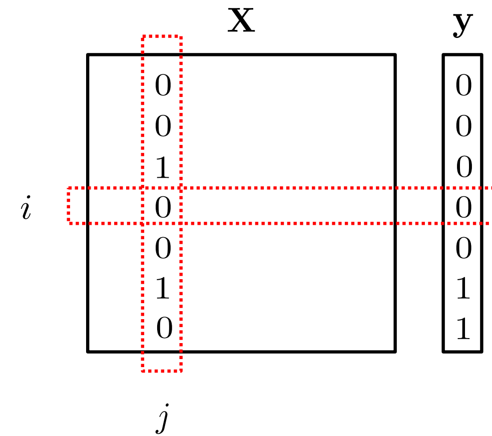
$$\alpha_{j,l,k} \stackrel{est}{=} \frac{\#\text{times } X_j = l \wedge Y = k}{\#\text{times } Y = k} = \frac{m_{j,l,k}}{n_k}$$

$$\alpha_k \stackrel{est}{=} \frac{\#\text{times } Y = k}{\text{data set size}} = \frac{n_k}{n}$$

Naive Bayes Classification

Assume: discrete $\mathcal{X}_j$, discrete $\mathcal{Y}$

generalized Bernoulli distribution



$$p(x_1, x_2, ..., x_d | y) = \prod_{j=1}^{d} p(x_j | y)$$

$$\alpha_{j,l,k} \stackrel{est}{=} \frac{m_{j,l,k} + \ell}{n_k + \ell |\mathcal{X}_j|}$$

$$\alpha_k \stackrel{est}{=} \frac{n_k + \ell}{n + \ell |\mathcal{Y}|}$$

ℓ = user-specified constant

$\ell = 1$ gives Laplace smoothing

$|\mathcal{X}_j|$ = the number of possible values of feature j

Naive Bayes Classification

$$\mathbf{X} \qquad \mathbf{y}$$

$$\begin{array}{|c|}\hline x_{1,j} \\ x_{2,j} \\ x_{3,j} \\ x_{4,j} \\ x_{5,j} \\ x_{6,j} \\ x_{7,j} \\ \hline\end{array} \quad \begin{array}{|c|}\hline 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ \hline\end{array}$$

Assume: continuous $\mathcal{X}_j$, discrete $\mathcal{Y}$

Gaussian distribution

$$\forall i \quad x_{i,j} \in \mathbb{R}$$

$$y_i \in \mathcal{Y}$$

$$\mu_{j,k} = \mathbb{E}[X_j|Y = k]$$

$$\sigma^2_{j,k} = \mathbb{V}[X_j|Y = k]$$

$$P(Y = k) = \alpha_k$$

How many parameters?

Naive Bayes Regression

Assume: continuous $\mathcal{X}_j$, continuous $\mathcal{Y}$

Gaussian distribution

$\forall i \quad x_{i,j} \in \mathbb{R}$

$y_i \in \mathbb{R}$

$$
\begin{array}{cc}
\mathbf{X} & \mathbf{y} \\
\begin{bmatrix} x_{1,j} \\ x_{2,j} \\ x_{3,j} \\ x_{4,j} \\ x_{5,j} \\ x_{6,j} \\ x_{7,j} \end{bmatrix} &
\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{bmatrix}
\end{array}
$$

$$p(x_1, x_2, ..., x_d | y) = \prod_{j=1}^{d} p(x_j | y)$$

$$p(x_j | y) = \frac{p(x_j, y)}{p(y)}$$ where $p(x_j, y)$ is 2D Gaussian

What if features are discrete?

Example: Naive Bayes Classifier w/ Redundant Features

Let A, B, and C be binary features, such that $B = C$. Let $\mathcal{Y} = \{-, +\}$

Let $P(-) = P(+) = \frac{1}{2}$. Let $P(A) = P(B) = P(C) = \frac{1}{2}$.

$$\Rightarrow \begin{array}{l} P(+|A) = P(A|+) \\ P(+|B) = P(B|+) \end{array}$$

Optimal decision

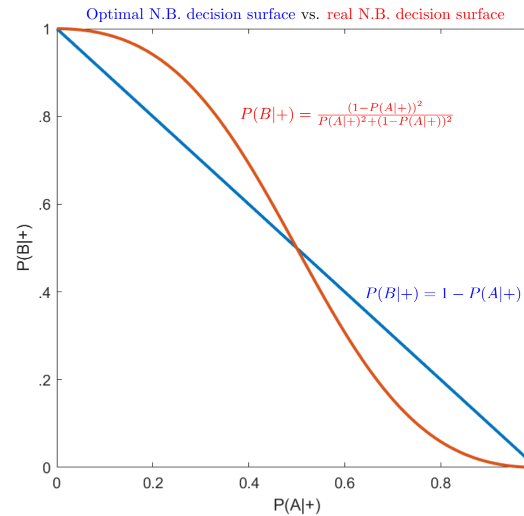
$$P(+|A, B, C) > P(-|A, B, C)$$

Naive Bayes optimal decision (C is ignored)

$$P(A|+)P(B|+) > P(A|-)P(B|-)$$

Naive Bayes decision

$$P(A|+)P(B|+)^2 > P(A|-)P(B|-)^2$$



Optimal Bayes Model

Given: a set of observations $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, $x_i \in \mathcal{X}, y_i \in \mathcal{Y}$

Objective: learn the posterior $p(y|x, \mathcal{D})$

Optimal Bayes Model:

$$p(y|x,\mathcal{D}) = \sum_{f \in \mathcal{F}} p(y|x, \mathcal{D}, f) p(f|x, \mathcal{D}) \qquad \text{Finite } \mathcal{F}$$

$$= \sum_{f \in \mathcal{F}} p(y|x, f) p(f|\mathcal{D})$$

Example: Let f_1, f_2 and f_3 be binary classifiers. Let $p(f_i|\mathcal{D}) = \{0.4, 0.3, 0.3\}$ and $P(Y=1|f_i) = \{1, 0, 0\}$.

What is the MAP prediction? What is the optimal Bayes prediction?