Summary Optimal Bayesian Classifier I

Ingredients:

1. **Stochastic variables** (X, Y) **with distribution** $P(\cdot, \cdot)$
2. **Loss function** $L(\cdot, \cdot)$
3. **Construct** $\hat{Y}(\cdot)$ **by minimizing** $EL(Y, \hat{Y}(X))$

Summary Optimal Bayesian Classifier I

Ingredients:

1. **Stochastic variables (X, Y) with distribution $P(\cdot, \cdot)$**

2. **Loss function $L(\cdot, \cdot)$**

3. **Construct $\hat{Y}(\cdot)$ by minimizing $EL(Y, \hat{Y}(X))$**

Solution:

$\forall x$, **solve** $\min_{\hat{Y}(x)} E_{Y|X=x} L(Y, \hat{Y}(x))$.

Summary Optimal Bayesian Classifier I

Ingredients:

1. **Stochastic variables** (X, Y) **with distribution** $P(\cdot,\cdot)$
2. **Loss function** $L(\cdot,\cdot)$
3. **Construct** $\hat{Y}(\cdot)$ **by minimizing** $EL(Y, \hat{Y}(X))$

Solution:

$\forall x$, **solve** $\min_{\hat{Y}(x)} E_{Y|X=x} L(Y, \hat{Y}(x))$.

In the binary case:

$$\hat{Y}(x) = I(\frac{P(Y=1|X=x)L(0,1)}{P(Y=0|X=x)L(1,0)} > 1) \tag{1}$$

This is of the form:

$$\hat{Y}(x) = I(\frac{P(Y=1|X=x)}{P(Y=0|X=x)} > c), \qquad c = \frac{L(1,0)}{L(0,1)} \tag{2}$$

We derived it last time when X **is a discrete s.v; the results are true in general.**

Observe that by Bayes rule:

$$P(Y = y|X = x) = P(X = x|Y = y)P(Y = y)\frac{1}{P(X = x)}$$

Observe that by Bayes rule:

$$P(Y = y|X = x) = P(X = x|Y = y)P(Y = y)\frac{1}{P(X = x)}$$

Hence

$$\hat{Y}(x) = I(\frac{P(Y = 1|X = x)}{P(Y = 0|X = x)} > c), \qquad c = \frac{L(1,0)}{L(0,1)} \tag{3}$$

Observe that by Bayes rule:

$$P(Y=y|X=x) = P(X=x|Y=y)P(Y=y)\frac{1}{P(X=x)}$$

Hence

$$\hat{Y}(x) = I(\frac{P(Y=1|X=x)}{P(Y=0|X=x)} > c), \qquad c = \frac{L(1,0)}{L(0,1)} \tag{4}$$

can be rewritten as:

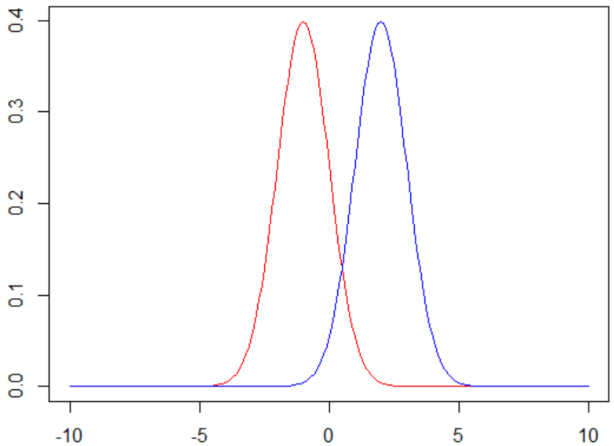
$$\hat{Y}(x) = I(\frac{P(X=x|Y=1)P(Y=1)}{P(X=x|Y=0)P(Y=0)} > c), \qquad c = \frac{L(1,0)}{L(0,1)} \tag{5}$$

$$\hat{Y}(x) = I(\frac{P(X=x|Y=1)}{P(X=x|Y=0)} > c_1), \qquad c_1 = \frac{L(1,0)P(Y=0)}{L(0,1)P(Y=1)} \tag{6}$$

Remember:

$$\hat{Y}(x) = I\left(\frac{P(X=x|Y=1)}{P(X=x|Y=0)} > c_1\right), \qquad c_1 = \frac{L(1,0)P(Y=0)}{L(0,1)P(Y=1)} \tag{7}$$

Suppose $X|Y=y \sim \mathcal{N}(\mu_y, \sigma^2)$



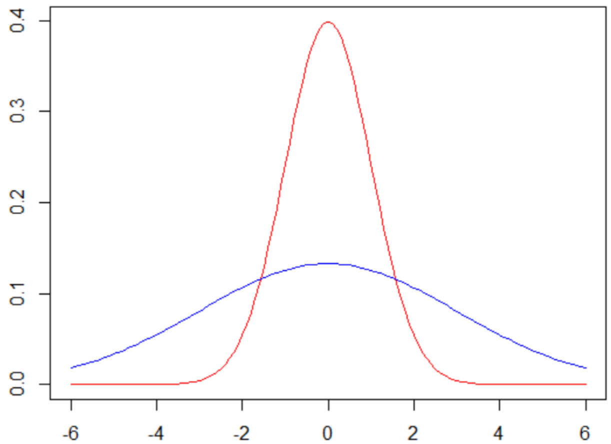
$P(X|Y=0)$, $P(X|Y=1)$,

How does $\hat{Y}(x)$ **look like?**

Remember:

$$\hat{Y}(x) = I\left(\frac{P(X = x|Y = 1)}{P(X = x|Y = 0)} > c_1\right), \qquad c_1 = \frac{L(1,0)P(Y = 0)}{L(0,1)P(Y = 1)} \tag{8}$$

Suppose $X|Y = y \sim \mathcal{N}(\mu, \sigma_y^2)$



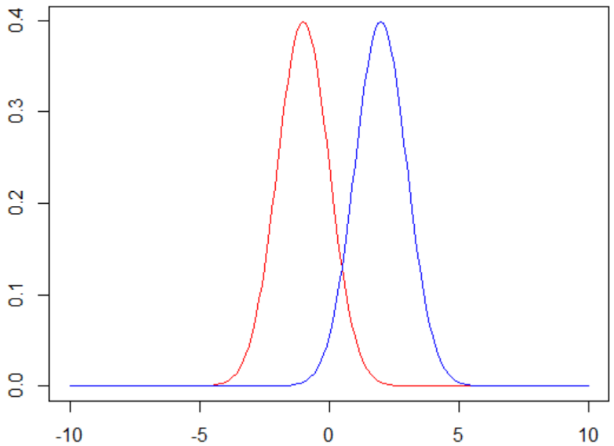
$\color{red}{P(X|Y=0)}$, $\color{blue}{P(X|Y=1)}$,

How does $\hat{Y}(x)$ **look like?**

Remember:

$$\hat{Y}(x) = I(\frac{P(X=x|Y=1)}{P(X=x|Y=0)} > c_1), \qquad c_1 = \frac{L(1,0)P(Y=0)}{L(0,1)P(Y=1)} \tag{9}$$

Suppose $X|Y=y \sim \mathcal{N}(\mu_y, \sigma^2)$



$\color{red}{P(X|Y=0)}$, $\color{blue}{P(X|Y=1)}$,

How does $\hat{Y}(x)$ **look like?**

Observe:

Strictly speaking not necesarry to know the (conditional) distributions, only their ratio.

We solve a harder problem than the original one.

Naive Bayesian Classifier

Remember:

$$\hat{Y}(x) = I\left(\frac{P(X=x|Y=1)}{P(X=x|Y=0)} > c_1\right), \qquad c_1 = \frac{L(1,0)P(Y=0)}{L(0,1)P(Y=1)} \tag{10}$$

How to define (estimate) $P(X=x|Y=y)$ in general?

This is not obvious if X is a (high dimensional) vector.

Naive Bayesian Classifier

Remember:

$$\hat{Y}(x) = I(\frac{P(X = x|Y = 1)}{P(X = x|Y = 0)} > c_1), \qquad c_1 = \frac{L(1,0)P(Y = 0)}{L(0,1)P(Y = 1)} \tag{11}$$

How to define (estimate) $P(X = x|Y = y)$ in general?

This is not obvious if X is a (high dimensional) vector.

The naive Bayesian classifier is based on the simplification (assumption) that $X|Y = y$ are independent s.v. : $P(X = x|Y = y) = \prod_i P(X_i = x_i|Y = y)$.

Instead of 1 d-dimensional problem, we have d 1-dimensional problems.

Reminder: multivariate normal distribution

We say $X = (X_1, \cdots, X_d) \sim \mathcal{N}(\mu, \Sigma)$ **if**

$$f_X(x) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp\left(-\frac{(x-\mu)^t \Sigma^{-1}(x-\mu)}{2}\right)$$

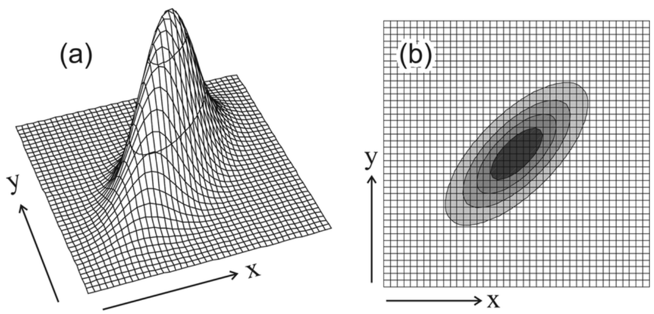
Reminder: multivariate normal distribution

We say $X = (X_1, \cdots, X_d) \sim \mathcal{N}(\mu, \Sigma)$ **if**

$$f_X(x) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp(-\frac{(x-\mu)^t \Sigma^{-1}(x-\mu)}{2})$$

Example:

Take $d = 2$**.**



(from: Keith Sircombe, axes shoud be x_1 **and** x_2**)**

Contours are ellipses (ellipsoides)

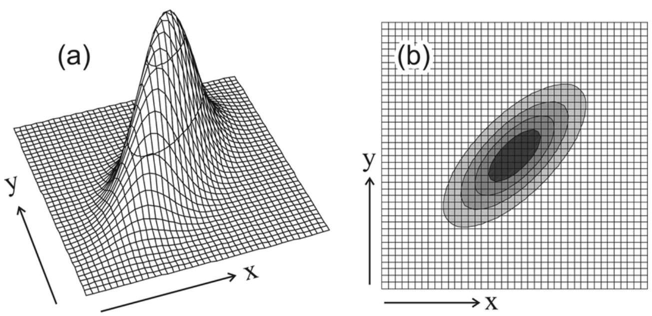
Reminder: multivariate normal distribution

We say $X = (X_1, \cdots, X_d) \sim \mathcal{N}(\mu, \Sigma)$ **if**

$$f_X(x) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp(-\frac{(x-\mu)^t \Sigma^{-1} (x-\mu)}{2})$$

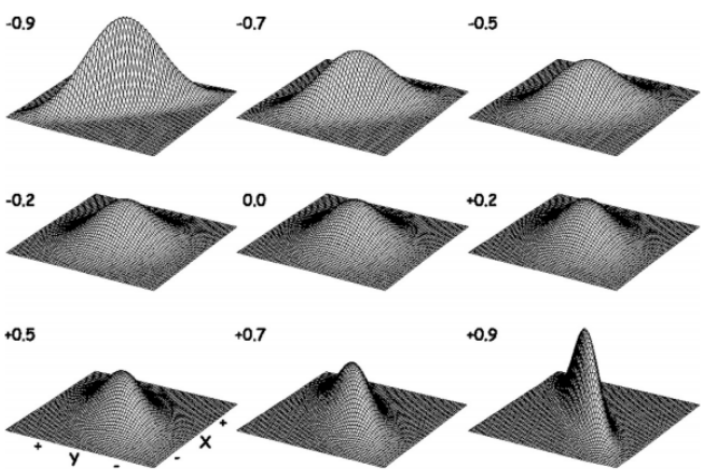
Example:

Take $d = 2$**.**



(from: Keith Sircombe)

Contours are ellipses (ellipsoides)



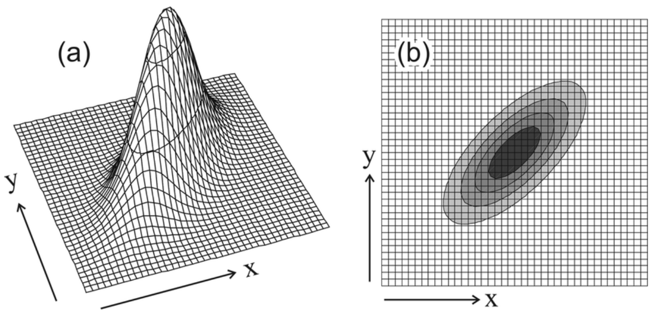
Reminder: multivariate normal distribution

We say $X \sim \mathcal{N}(\mu, \Sigma)$ if

$$f_X(x) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp\left(-\frac{-(x-\mu)^t\Sigma^{-1}(x-\mu)}{2}\right)$$

Example:

Take $d = 2$.



(from: Keith Sircombe)

Contours are ellipses (ellipsoides)

Properties:

- $EX = \mu$
- $Cov(X) := [Cov(X_i, X_j)]_{i,j} = \Sigma$

Properties:

- **$l^t X$ is also normal distributed. In general $Y = \mathbb{A}X \sim \mathcal{N}(\mathbb{A}\mu, \mathbb{A}\Sigma\mathbb{A}^t)$**
- **Marginal and conditional distributions are also normal.**
- **X has independent components if and only if Σ is diagonal matrix.**

Optimal Bayesian Classifier II

Remember:

$$\hat{Y}(x) = I\left(\frac{P(X = x|Y = 1)}{P(X = x|Y = 0)} > c_1\right), \qquad c_1 = \frac{L(1,0)P(Y = 0)}{L(0,1)P(Y = 1)} \tag{12}$$

Case 1: $X|Y = y \sim \mathcal{N}(\mu_y, \Sigma)$

How does $\hat{Y}(x)$ look like?

Optimal Bayesian Classifier II

Remember:

$$\hat{Y}(x) = I(\frac{P(X = x|Y = 1)}{P(X = x|Y = 0)} > c_1), \qquad c_1 = \frac{L(1,0)P(Y = 0)}{L(0,1)P(Y = 1)} \tag{13}$$

Case 1: $X|Y = y \sim \mathcal{N}(\mu_y, \Sigma)$

How does $\hat{Y}(x)$ look like?

Remember

$$f_X(x) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp(-\frac{(x-\mu)^t\Sigma^{-1}(x-\mu)}{2})$$

$$log\frac{P(X = x|Y = 1)}{P(X = x|Y = 0)} = \frac{-(x-\mu_1)^t\Sigma^{-1}(x-\mu_1)}{2} + \frac{(x-\mu_0)^t\Sigma^{-1}(x-\mu_0)}{2} + c^*$$

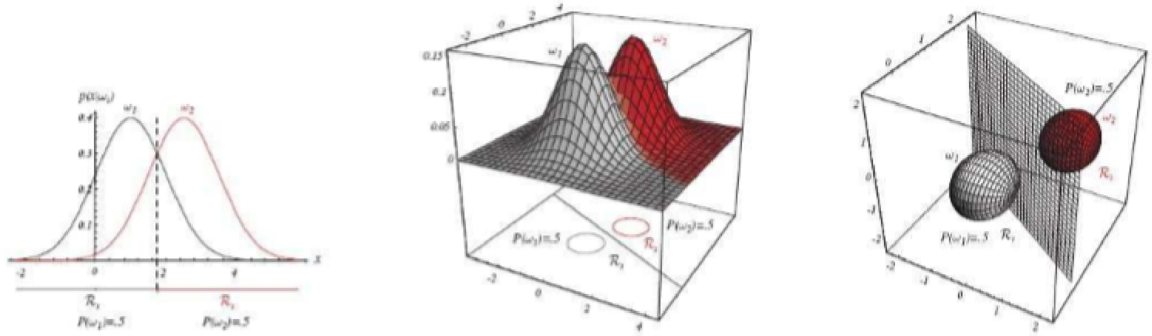
RHS is of the form

$$\Sigma^{-1}(\mu_1 - \mu_0)x + c^{**}$$

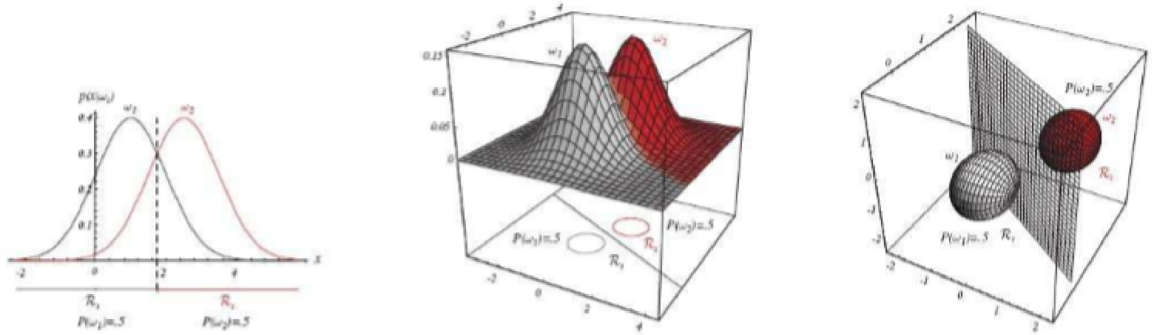
Hence classifier is of the form:

$$\hat{Y}(x) = I(l^t x > c^{***}), \qquad l = \Sigma^{-1}(\mu_1 - \mu_0) \tag{14}$$

This is called the Linear Discriminant Analysis Classifier (LDA)



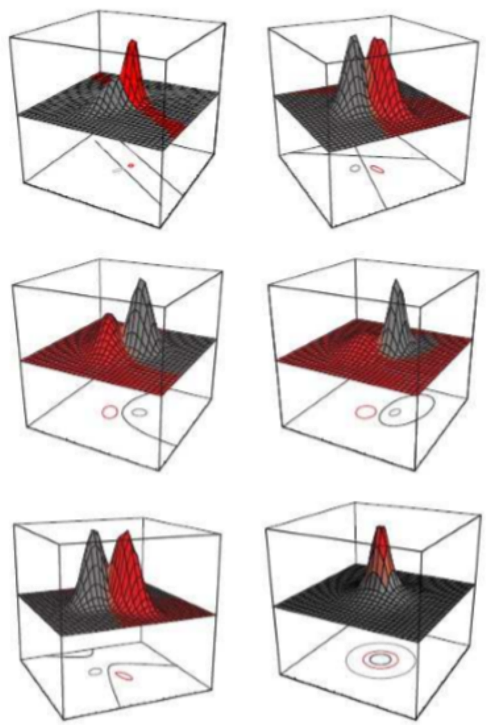
(from: Duda and Hart book)



(from: Duda and Hart book)

Case 2: $X|Y = y \sim \mathcal{N}(\mu_y, \Sigma_y)$

How does $\hat{Y}(x)$ look like?



(from: Duda and Hart book)

This is called the Quadratic Discriminant Analysis Classifier (QDA)

k- Nearest Neighbor classifier

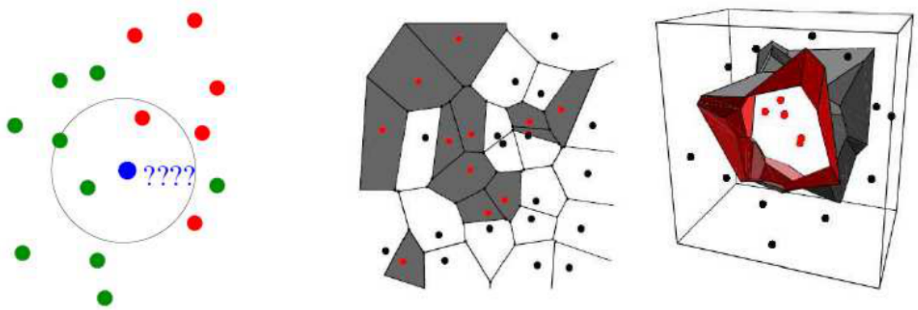
Idea:

Given x. To decide $\hat{y}(x)$, look for $N_k(x)$ the set of the k nearest observations to x.

For clasification: **decide by voting: assign most frequent category in $\{y_i, i \in N_k(x)\}$**

For regression: **decide by averaging: calculate mean of $\{y_i, i \in N_k(x)\}$**

Example for classification:



(LHS from: Duda and Hart book)

For $k = 1$: relation with Voronoi diagram

k- Nearest Neighbor classifier

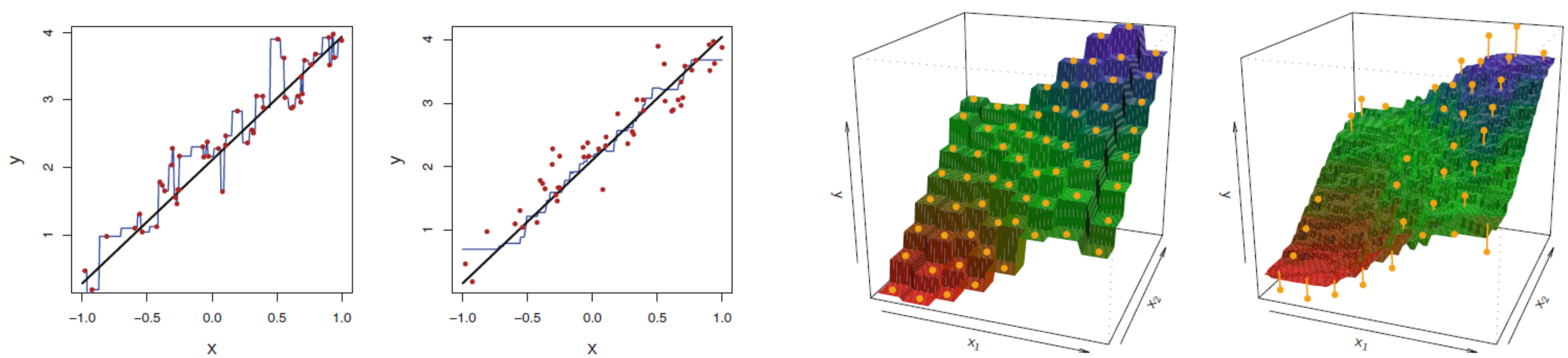
Idea:

Given x. To decide $\hat{y}(x)$, look for $N_k(x)$ the set of the k nearest observations to x.

For clasification: **decide by voting: assign most frequent category in $\{y_i, i \in N_k(x)\}$**

For regression: **decide by averaging: calculate mean of $\{y_i, i \in N_k(x)\}$**

Example for regression:



(from: ISL book)

$k = 1$ versus $k = 7$

Quiz: What properties do you observe?

Denote by L^* the (mean) error of the optimal Bayes classifier, f_n a classifier based on $\{(X_i, Y_i)\}_1^n$, and $L(f_n) = E(L(Y, f_n(X))$, one can prove:

Property 1 **If $n \to \infty$ and f_n is 1-NN classifier:**

$$L^* \leq EL(f_n) \leq 2 * L^*$$

Property 2 **If $n \to \infty$ and $k \to \infty$ such that $k/n \to 0$, if f_n is k-NN classifier: for any $P()$:**

$$EL(f_n) \to L^*$$

On the other hand:

Theorem no free lunch: **for finite n, within any assumption about P, no classifier is superior.**