

# Personalized medicine: Individualized Treatment Rules

# Definition of Precision Medicine

- Precision medicine is a medical model that proposes the customization of healthcare, with medical decisions, practices, and/or products being tailored to the individual patient. In this model, diagnostic testing is often employed for selecting appropriate and optimal therapies based on the context of a patient's genetic content or other molecular or cellular analysis. Tools employed in precision medicine can include molecular diagnostics, imaging, and analytics/software.
- Making optimal healthcare decision for each individual patient based on this subject's context information.

# Illustration Data

Table: An illustration dataset

ID	$Y$	$A$	$X_1$	$X_2$	$X_3$	$\dots$
1	1.5	1	F	26	7.8	$\dots$
2	1.2	2	M	28	8.2	$\dots$
3	2.3	3	M	31	8.9	$\dots$
4	0.9	2	F	35	9.4	$\dots$
5	1.7	1	M	22	7.3	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

## Research Question

Based on these data, how can we treat a new patient?

In other words, how can we learn a treatment assignment rule that, if followed by the entire population of certain patients, would lead to the best outcome on average?

# Three Key Components for Precision Medicine

Context based decision learning has data in 3 components:

- $X_1, X_2, \dots, X_p$  is context information.
- $A$  is a context action.
- $Y$  is a reward.

Notes:

- This data structure differs from data for typical supervised and unsupervised learning.
- Examples on data collection for precision medicine ...

## Other Examples: Car Purchase

Table: My Friends' Rating of Their First Cars

ID	Satisfaction	Car Type	Gender	Age	Mileage per Day	...
1	90%	Focus	F	26	7.8	...
2	85%	Corolla	M	28	8.2	...
3	70%	Civic	M	31	8.9	...
4	75%	Corolla	F	35	9.4	...
5	60%	Civic	M	22	7.3	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Learning from these data, what car should I purchase?

## Other Examples: Business Investment

**Table:** Previous Commercial Investments and Returns

Case ID	Return	Type	Month	Location	Share of Market	...
1	1.2	TV	Jan	MW	12.5	...
2	0.9	Radio	Oct	NE	18.2	...
3	1.4	Web	Nov	WE	12.9	...
4	1.3	Web	Dec	MW	10.4	...
5	1.2	Radio	Feb	SE	11.3	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Learning from these data, what is our best way to invest in New England area if our product has 12% market share in this March?

# Making Optimal Decision Based on Data

Broad applications, some examples:

- Treatment selection: which treatment is the best for this patient?
- Treatment transition: should we keep using the current treatment or consider an intensification?
- Business analytic: how to invest (among a few choices) to maximize the return?
- Recommendation system: which item should a system recommend to a customer to maximize profit?

All these problems are similar in terms of data format and analytic solutions. **Essentially, we focus on a problem of making the optimal decision based on data.**

So, what is a general framework to solve this?

# Reinforcement Learning Framework

- This problem is a special case in reinforcement learning framework which is different from supervised learning (e.g. classification) and unsupervised learning (e.g. clustering).
- Traditional alternatives (e.g. linear regression for rewards) may not be efficient to solve these problems.
- It is connected with supervised learning methods (e.g. support vector machines).
- It can be extended to multiple stage decision making to optimize treatment sequences (e.g. dynamic treatment regimes).



# Outlines

- 1 Precision Medicine
- 2 Support Vector Machines and Outcome Weighted Learning
- 3 Extensions: multi-treatments, ordinal treatments

# Notations

- There are  $N$  subjects from a large population.
- $A_i$  is the treatment assignment (actions), where  $i = 1, \dots, N$ .
- $Y_i$  is the response assuming that larger  $Y_i$  is better (rewards).
- $X_i$  is a vector of covariates.
- $(Y, A, X)$  is the generic random variable of  $\{(Y_i, A_i, X_i)\}$ .
- $\mathcal{P}$  is the distribution of  $(Y, A, X)$ .
- $E$  is the expectation with respect to  $\mathcal{P}$ .
- Population space  $\mathcal{X}$ , i.e.  $X_i \in \mathcal{X}$ .
- $\mathcal{D}(\cdot)$  is a treatment recommendation based on covariates, i.e.  $\mathcal{D}(\cdot) : \mathcal{X} \rightarrow \mathcal{A}$ .
- $\mathcal{P}^{\mathcal{D}}$  is the distribution of  $(Y, A, X)$  given that  $A = \mathcal{D}(X)$ .

# Value Function

Define

$$E^{\mathcal{D}}(Y) = \int Y d\mathcal{P}^{\mathcal{D}} = \int Y \frac{d\mathcal{P}^{\mathcal{D}}}{d\mathcal{P}} d\mathcal{P} = E \left[ \frac{I\{A = \mathcal{D}(X)\}}{p(A|X)} Y \right],$$

where we use the fact that

$$\frac{d\mathcal{P}^{\mathcal{D}}}{d\mathcal{P}} = \frac{p(y|x, a)I\{a = \mathcal{D}(x)\}p(x)}{p(y|x, a)p(a|x)p(x)} = \frac{I\{a = \mathcal{D}(x)\}}{p(a|x)}.$$

Our objective is to find  $\mathcal{D}(\cdot)$  to maximize the following **value function**:

$$\mathcal{D}_o \in \operatorname{argmax}_{\mathcal{D} \in R} E^{\mathcal{D}}(Y) = E \left[ \frac{I\{A = \mathcal{D}(X)\}}{p(A|X)} Y \right], \quad (1)$$

where  $R$  is a space of possible treatment recommendations.

# Advantages of This Framework

- $Y$  is able to handle binary, continuous, time to event data type.
- $A$  is able to handle multiple treatments.
- $X$  is able to incorporate variety of variables. For example, if  $X$  includes study ID, the framework can be used for meta analysis.
- $P(A|X)$  allows treatment assignments depending on covariates. So it can handle both randomized control trials and observational studies.
- It has an objective function to evaluate different treatment assignments.

# An Example to Build Intuition

Table: Example Data

ID	$Y$	$A$	$X$	$P(A X)$
1	1	1	1	0.5
2	2	1	2	0.5
3	3	1	3	0.5
4	4	1	4	0.5
5	5	1	5	0.5
6	3	2	1	0.5
7	3	2	2	0.5
8	3	2	3	0.5
9	3	2	4	0.5
10	3	2	5	0.5

Questions to think about: why is  $P(A|X) = 0.5$ ? what do the responses look like?

# Which Doctor is Better

Suppose we have two doctors and each of them has a treatment rule.  
Which doctor is a better one?

- Doctor Adam: give patients treatment 1 if  $X \geq 2$ , and treatment 2 otherwise, denoted as  $\mathcal{D}_A(X)$ .
- Doctor Barry: give patients treatment 1 if  $X \geq 3$ , and treatment 2 otherwise, denoted as  $\mathcal{D}_B(X)$ .

## Example Continued

Table: Calculation Based on Table 4

ID	$Y$	$A$	$X$	$P(A X)$	$\mathcal{D}_A$	$\mathcal{D}_B$	$\mathcal{D}_A = A$	$\mathcal{D}_B = A$
1	1	1	1	0.5	2	2	0	0
2	2	1	2	0.5	1	2	1	0
3	3	1	3	0.5	1	1	1	1
4	4	1	4	0.5	1	1	1	1
5	5	1	5	0.5	1	1	1	1
6	3	2	1	0.5	2	2	1	1
7	3	2	2	0.5	1	2	0	1
8	3	2	3	0.5	1	1	0	0
9	3	2	4	0.5	1	1	0	0
10	3	2	5	0.5	1	1	0	0

## Example Continued

Doctor Adam:

$$\begin{aligned} E^{\mathcal{D}_A(Y)} &= \frac{1}{10} \left( \frac{0}{0.5} \times 1 + \frac{1}{0.5} \times 2 + \frac{1}{0.5} \times 3 + \frac{1}{0.5} \times 4 + \frac{1}{0.5} \times 5 + \frac{1}{0.5} \right. \\ &\quad \left. \times 3 + \frac{0}{0.5} \times 3 + \frac{0}{0.5} \times 3 + \frac{0}{0.5} \times 3 + \frac{0}{0.5} \times 3 \right) \\ &= 3.4 \end{aligned}$$

Doctor Barry:

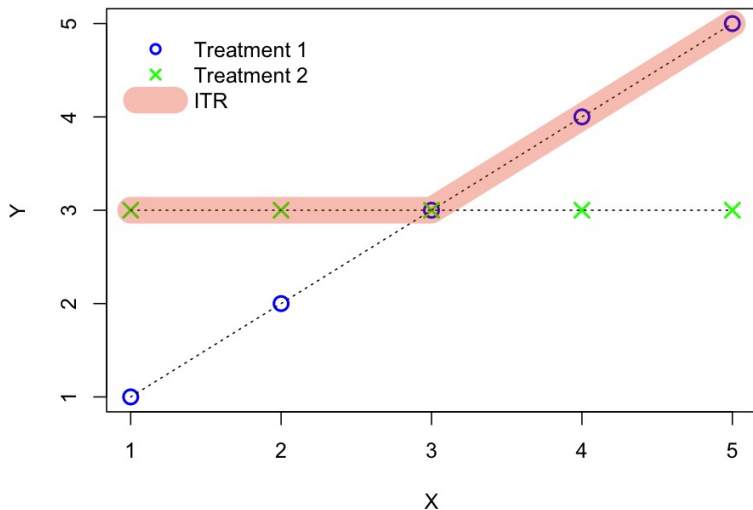
$$\begin{aligned} E^{\mathcal{D}_B(Y)} &= \frac{1}{10} \left( \frac{0}{0.5} \times 1 + \frac{0}{0.5} \times 2 + \frac{1}{0.5} \times 3 + \frac{1}{0.5} \times 4 + \frac{1}{0.5} \times 5 + \frac{1}{0.5} \right. \\ &\quad \left. \times 3 + \frac{1}{0.5} \times 3 + \frac{0}{0.5} \times 3 + \frac{0}{0.5} \times 3 + \frac{0}{0.5} \times 3 \right) \\ &= 3.6 \end{aligned}$$

Conclusion: Doctor Barry's rule is better than Doctor Adam's. Can we improve Doctor Barry's rule? How can we find the best rule?



# Graphic Illustration

## Individualized Treatment Recommendation



# Thought Provoking Questions

- Both treatment 1 and treatment 2 have an average treatment effect as 3.0. But ITR generates average benefit value 3.6. Can algorithm beat a new molecule entity?
- Treatment 1 should not be only better than treatment 2. It has to be better with a non-trivial benefit margin. How can we handle this case?
- What if the treatment randomization ratio is not 1:1?
- What if we have multiple covariates? The rule can be complicated.
- What if we have multiple treatments?

## Analysis results: how ITR creates more value.

This data analysis shows how ITR creates additional value for patients. We have 1978 patients from two treatment arms, and 2 important biomarkers are selected from 35 biomarkers.

**Table:** *HbA1c Reduction Before and After Following ITR.* Patients with baseline fasting insulin  $\geq 61.12\text{pmol/L}$  and baseline HbA1c  $\geq 8.1\%$  ( $A_o^1$ ) are recommended to take Pioglitazone, otherwise ( $A_o^0$ ) patients are recommended to take Gliclazide. After following ITR, the overall HbA1c reduction changes from -1.287% to -1.473%. Notes: ITR is our proposed method which is referred to as Individualized Treatment Recommendation.

Original			Follow ITR	
-1.287			-1.473	
	Gliclazide	Pioglitazone	Gliclazide	Pioglitazone
Mean	-1.271	-1.303	$A_o^1$ -1.394	-1.864
			$A_o^0$ -1.19	-0.932

# Key Insights on Solving ITR

Three connections:

- ➊ Maximization and minimization of the value function.
- ➋ Classification and loss functions.
- ➌ ITR and weighted classifications.

# From Maximization to Minimization

Original objective function (Qian and Murphy, 2011)

$$\mathcal{D}_o \in \operatorname{argmax}_{\mathcal{D} \in \mathcal{R}} E^{\mathcal{D}}(Y) = E \left[ \frac{I \{A = \mathcal{D}(X)\}}{p(A|X)} Y \right]. \quad (2)$$

Making connections:

$$E \left\{ \frac{Y}{p(A|X)} \right\} - E \left[ \frac{I \{A = \mathcal{D}(X)\}}{p(A|X)} Y \right] = E \left[ \frac{I \{A \neq \mathcal{D}(X)\}}{p(A|X)} Y \right]$$

New objective function:

$$\mathcal{D}_o \in \operatorname{argmin}_{\mathcal{D} \in \mathcal{R}} E^{\mathcal{D}}(Y) = E \left[ \frac{I \{A \neq \mathcal{D}(X)\}}{p(A|X)} Y \right]. \quad (3)$$

# Empirical Evaluation

Objective function:

$$\mathcal{D}_o \in \operatorname{argmin}_{\mathcal{D} \in \mathcal{R}} E^{\mathcal{D}}(Y) = E \left[ \frac{I \{A \neq \mathcal{D}(X)\}}{p(A|X)} Y \right].$$

When we have data, we can evaluate the objective function as,  
Empirical evaluation:

$$D_o = \operatorname{argmin}_{D \in \mathcal{R}} n^{-1} \sum_{i=1}^n \frac{Y_i}{p(A_i|X_i)} I \{A_i \neq \mathcal{D}(X_i)\}. \quad (4)$$

# Classification Problems

A classification problem is to train a rule  $\mathcal{D}(X)$  on a dataset to predict new subject membership. A simple dataset can be as below,

Table: An illustration dataset

ID	$A$	$X_1$	$X_2$	$X_3$	$\dots$
1	1	F	26	7.8	$\dots$
2	2	M	28	8.2	$\dots$
3	1	M	31	8.9	$\dots$
4	3	F	35	9.4	$\dots$
5	1	M	22	7.3	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

# Classification and Loss Function

Roughly speaking, A good classifier has smaller errors (we will add regularization later).

Classification objective function

$$D_o = \operatorname{argmin}_{D \in R} n^{-1} \sum_{i=1}^n I \{A_i \neq \mathcal{D}(X_i)\}.$$

Compare our ITR objective function below

ITR objective function

$$D_o = \operatorname{argmin}_{D \in R} n^{-1} \sum_{i=1}^n \frac{Y_i}{p(A_i|X_i)} I \{A_i \neq \mathcal{D}(X_i)\}.$$



# Important Implications and Next Steps

- We can solve the original reinforcement learning problem (ITR) as a weighted supervised learning problems (Zhao et al., 2012).
- There are vast amount of methods and literatures on solving classification problems, in particular for binary classifications.
- With some modifications, we can leverage these existing algorithms to develop our ITR algorithms.
- We will focus on the Support Vector Machine (SVM). However, many other classifiers can be tailored for this.

# Outlines

- 1 Precision Medicine
- 2 Support Vector Machines and Outcome Weighted Learning
- 3 Extensions: multi-treatments, ordinal treatments

# Classification

- Observe a collection of i.i.d. training data  $(X_1, a_1), (X_2, a_2), \dots, (X_n, a_n)$  from  $\mathcal{P}$ .
- Covariates (inputs, features, prediction variables):  
 $X_i = (X_{i1}, \dots, X_{ip})$
- Response variable (class label, output):

$$a_i \in \{c_1, c_2, \dots, c_K\}.$$

- We want to build a model  $\mathcal{D}(X)$  (using the training data), so that when seeing a new input vector  $X$ , we can predict the output  $\hat{a}$ .

# Classification Errors and Loss Function

- Loss function (0/1):

$$L\{A, \mathcal{D}(X)\} = \begin{cases} 0 & \text{if } A = \mathcal{D}(X) \\ 1 & \text{if } A \neq \mathcal{D}(X) \end{cases}$$

- Misclassification error

$$\begin{aligned} R(\mathcal{D}) &= E_{\mathcal{P}} L\{A, \mathcal{D}(X)\} \\ &= P_{\mathcal{P}}[I\{A \neq \mathcal{D}(X)\}]. \end{aligned}$$

- For binary class case, Bayes optimal classifier ( $A \in \{-1, 1\}$ ):

$$\begin{aligned} \mathcal{D}^*(X_i) &= \underset{\mathcal{D}}{\operatorname{argmin}} R(\mathcal{D}) \\ &= \operatorname{sgn} \{ \Pr(A = 1 | X = X_i) - \Pr(A = -1 | X = X_i) \}. \end{aligned}$$

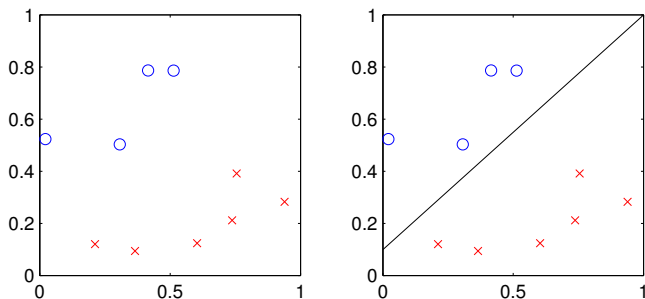
- Bayes error:  $R(\mathcal{D}^*)$ .

# Binary Large-Margin Classifier

- $a \in \{\pm 1\}$ ;  
Estimate  $f(X)$  with classification rule  $\text{sgn}\{f(X)\} : \mathbb{R}^d \rightarrow \{\pm 1\}$ ,  
 $\hat{a} = +1$  if  $f(X) \geq 0$  and  $\hat{y} = -1$  if  $f(X) < 0$ .
- $A_i f(X_i)$ : functional margin.
- Correction classification if  $A_i f(X_i) > 0$ .
- The 0–1 loss:  $I\{A_i f(X_i) \leq 0\}$ .

# Support Vector Machine (SVM)

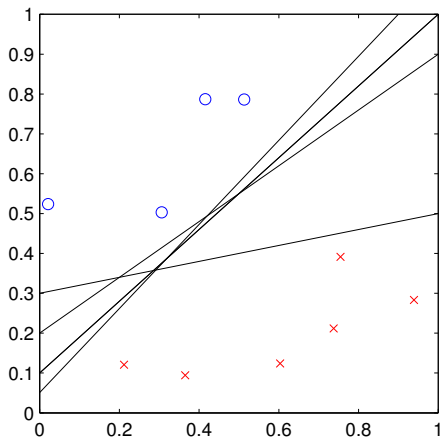
Linearly separable: Find  $f(X) = \beta_0 + X^\top \beta$  to separate two groups of points.



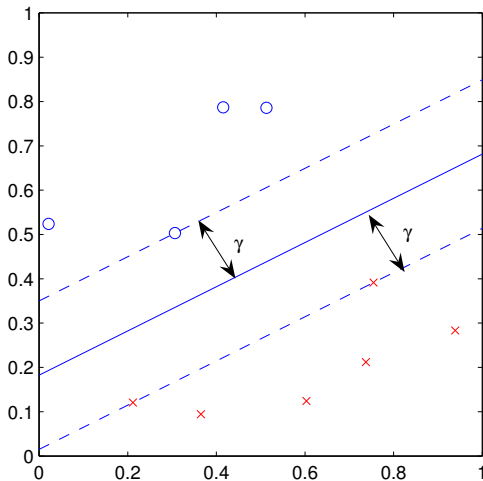
Note:

● red cross  $\longleftrightarrow +1$ ;    blue circle  $\longleftrightarrow -1$ .

Which one is the best?



# SVM: Maximum Separation





# Maximum Margin Classifier

**Goal:** Separate two classes and maximizes the distance to the closest points from either class (Vapnik 1996)

- Unique solutions
- Better classification performance on the training data

$$\begin{array}{ll} \underset{\beta, \beta_0}{\text{maximize}} & \gamma \\ \text{subject to} & A_i(\beta_0 + X_i^\top \beta) \geq \gamma, \\ & \|\beta\| = 1. \end{array}$$

All the points are at least a signed distance  $\gamma$  from the decision boundary

- Maximize the minimum distance
- Need constraint  $\|\beta\| = 1$

## Equivalent Problem

Try to get rid of the constraint  $\|\beta\|=1$

$$\frac{1}{\|\beta\|} A_i(X_i^\top \beta + \beta_0) \geq \gamma,$$

or equivalently

$$A_i(X_i^\top \beta + \beta_0) \geq \gamma \|\beta\|$$

Any positively scaled  $(\beta, \beta_0)$  also satisfies this inequality. We set  $\|\beta\| = \frac{1}{\gamma}$ . Then the objective function  $\gamma = 1/\|\beta\|$ , and

$$\begin{aligned} & \underset{\beta, \beta_0}{\text{minimize}} && \frac{1}{2} \|\beta\| \\ & \text{subject to} && A_i(X_i^\top \beta + \beta_0) \geq 1, \quad \forall i = 1, \dots, n. \end{aligned}$$

Linear SVM for perfectly separable cases.

Note: by definition  $1/\|\beta\|$  is the width of margin.

# Optimal Hyperplane of SVM

$$\begin{array}{ll}\text{minimize} & \frac{1}{2}\|\beta\|^2 \\ \text{subject to} & A_i(\langle\beta, X_i\rangle + \beta_0) - 1 \geq 0, \quad \forall i = 1, 2, \dots, n.\end{array}$$

- Lagrange function is :

$$L_P(\beta, \beta_0, \alpha) = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^n \alpha_i \{A_i(\langle\beta, X_i\rangle + \beta_0) - 1\}$$

- For any fixed  $\alpha$ :

$$\left\{ \begin{array}{l} \frac{\partial L(\beta, \beta_0, \alpha)}{\partial \beta_j} = 0, \\ \frac{\partial L(\beta, \beta_0, \alpha)}{\partial \beta_0} = 0 \end{array} \right. \quad j = 1, 2, \dots, p \quad \Longrightarrow \quad \left\{ \begin{array}{l} \beta = \sum_{i=1}^n \alpha_i A_i X_i \\ 0 = \sum_{i=1}^n \alpha_i A_i \end{array} \right.$$

# The Dual Problem

$$\begin{aligned} \text{maximize} \quad & L_D(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j A_i A_j \langle X_i, X_j \rangle \\ \text{subject to} \quad & \alpha_i \geq 0, \quad i = 1, 2, \dots, n \\ & \sum_{i=1}^n \alpha_i A_i = 0. \end{aligned}$$

This optimization is a **quadratic programming problem** and can be solved using classical optimization software. We are going to provide details on implementation in the R hands on example.

# Primal vs. Dual

- Minimize  $L_P$  with respect to primal variables  $\beta_0, \beta$
- Maximize  $L_D$  with respect to dual variables  $\alpha_i$
- Maximizing the dual is often a simpler convex QP than the primal, in particular when  $p \gg n$ .

# Recovering the Optimal Hyperplane

- The optimizer of the dual:  $\alpha^*$
- $\beta^*$  is given by:

$$\beta^* = \sum_{i=1}^n \alpha_i^* A_i X_i.$$

- $\beta_0^*$  ???
- Decision function:

$$f(\mathbf{x}) = \langle \beta^*, X \rangle + \beta_0^*.$$

- Classification rule:

$$\text{sgn}\{f(X)\}.$$

# Support Vectors

The KKT conditions imply,

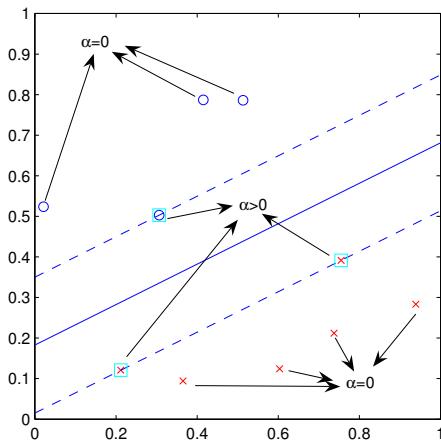
$$\alpha_i^* \left\{ A_i(\beta_0^* + X_i^\top \beta^*) - 1 \right\} = 0.$$

These imply

- If  $A_i f^*(X_i) > 1$ , then  $\alpha_i^* = 0$ .
- If  $\alpha_i^* > 0$ , then  $A_i f^*(X_i) = 1$ , or in other words,  $X_i$  is on the boundary of the “slab”.
- The solution  $\beta^*$  is defined in terms of a linear combination of the support points.

# Geometric Interpretation: Support Vectors

The  $i$ -th point is called a support vector if  $\alpha_i > 0$



The  $i$ -th point is a support vector  $\implies A_i(\langle \beta^*, X_i \rangle + \beta_0) = 1 \implies \beta_0^* = \dots$



# General Case for SVM

- Nonseparable: “zero”-error not attainable  $\rightarrow$  “slack variables”  $\{\xi_i\}_{i=1}^n$

$$\underset{\beta, \beta_0, \xi}{\text{minimize}} \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^n \xi_i$$

$$\begin{aligned} \text{subject to } & A_i f(X_i) \geq (1 - \xi_i), \quad i = 1, \dots, n, \\ & \xi_i \geq 0, \quad i = 1, \dots, n, \end{aligned}$$

where  $C > 0$  is a tuning parameter.

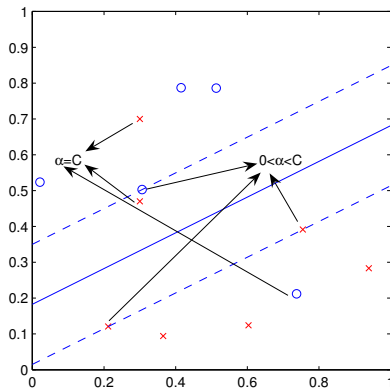
# The Dual Problem for SVM

Substituting into the Lagrange primal, we obtain the Lagrange dual problem as

$$\begin{array}{ll}\text{minimize} & L_D(\alpha) = \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j A_i A_j \langle X_i, X_j \rangle - \sum_{i=1}^n \alpha_i \\ \text{subject to} & 0 \leq \alpha_i \leq C, \quad i = 1, 2, \dots, n \\ & \sum_{i=1}^n \alpha_i A_i = 0.\end{array}$$

- Can be solved by quadratic programming.
- Recover  $\beta$ :  $\beta = \sum_{i=1}^n \alpha_i A_i X_i$ ; For given  $\beta$ ,  $\beta_0$  can be solved using KKT conditions or Linear Programming (LP).

# Support Vectors



- $\alpha_i = 0 \rightarrow A_i f(X_i) > 1$ ; not needed in constructing  $f(X)$ .

Support vectors:

- $0 < \alpha_i < C \rightarrow A_i f(X_i) = 1$  (Solve  $\beta_0$ ).
- $\alpha_i = C \rightarrow A_i f(X_i) < 1$ .
- **Outliers are SVs!**

# Reformulation of SVM Optimization

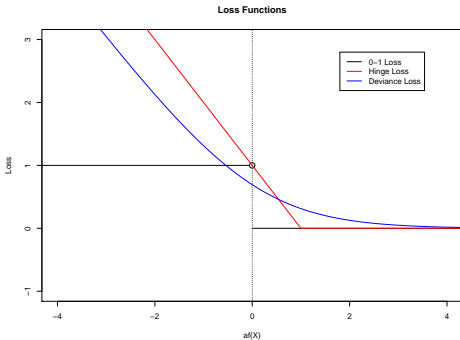
- SVM solves

$$\underset{\beta_0, \beta_1}{\text{minimize}} \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^n \ell_{\text{SVM}}\{A_i f(X_i)\}.$$

- $\ell_{\text{SVM}}(u) = (1 - u)_+$  (Hinge Loss).
- Nonlinear learning can be achieved by basis expansion or kernel learning.
- Kernel Trick: Replace  $\langle X_i, X_j \rangle$  by  $K(X_i, X_j)$  and  $f(X) = \sum_{i=1}^n A_i \alpha_i K(X_i, X) + \beta_0$ .

# Loss Functions

To estimate the classifier (threshold),  
 $\text{sgn}\{\Pr(A = 1|X) - \Pr(A = -1|X)\}$



- **0-1 Loss:**  
 $\ell\{A, f(X)\} = I\{Af(X) < 0\}.$
- **Hinge Loss:**  
 $\ell\{A, f(X)\} = \{1 - Af(X)\}_+$
- **Deviance Loss:**  $\ell\{A, f(X)\} = \log[1 + \exp\{-Af(X)\}]$

# Classification and Loss Function

Roughly speaking, a good classifier has smaller errors (will add regularization later).

Classification objective function

$$D_o = \operatorname{argmin}_{D \in R} n^{-1} \sum_{i=1}^n I \{A_i \neq \mathcal{D}(X_i)\}.$$

If we compare our ITR objective function as below,

ITR objective function

$$D_o = \operatorname{argmin}_{D \in R} n^{-1} \sum_{i=1}^n \frac{Y_i}{p(A_i|X_i)} I \{A_i \neq \mathcal{D}(X_i)\}.$$

# ITR.SVM

## ITR objective function

$$D_o = \operatorname{argmin}_{D \in \mathcal{R}} n^{-1} \sum_{i=1}^n \frac{Y_i}{p(A_i|X_i)} I\{A_i \neq \mathcal{D}(X_i)\}.$$

## ITR.SVM

$$D_o = \operatorname{argmin}_{D \in \mathcal{R}} n^{-1} \sum_{i=1}^n \frac{Y_i}{p(A_i|X_i)} \ell\{A_i, f(X_i)\} + \frac{\lambda}{2} \|f\|_{\mathcal{H}_K}^2.$$

R package: DTRlearn (function: Olearning\_Single)

# Outlines

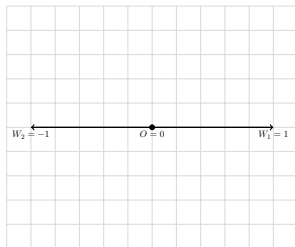
- 1 Precision Medicine
- 2 Support Vector Machines and Outcome Weighted Learning
- 3 Extensions: multi-treatments, ordinal treatments



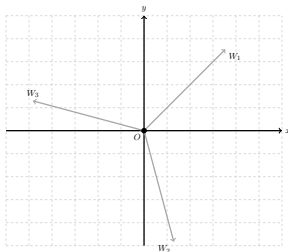
# Multicategory Angle-Based Classification (ABC)

- A simplex based classification structure
- Advantages of ABC (Zhang and Liu, Biometrika, 2014)
  - ▶ General structure: binary  $\rightarrow$  multicategory
  - ▶ Clear geometric interpretation
  - ▶ Free of sum-to-zero constraint  $\Rightarrow$  faster computational speed
  - ▶ Theoretical advantages
  - ▶ Numerically competitive

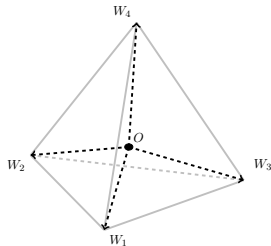
## Illustration of $\{W_i\}$ When $k = 2, 3, 4$ .



(a)  $k = 2$



(b)  $k = 3$



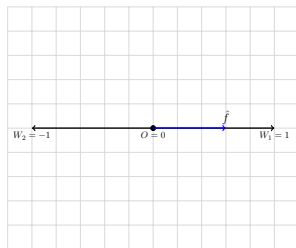
(c)  $k = 4$

Remark: When  $k = 3$ ,  $\{W_i, i = 1, 2, 3\}$  are the vertices of an equilateral triangle, and when  $k = 4$ ,  $\{W_i, i = 1, 2, 3, 4\}$  are the vertices of a regular tetrahedron.

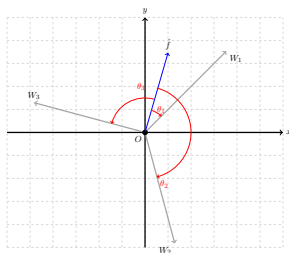
# Angle Based Classifier

- Let  $W_j$  represent class  $j$ .
- Our method is to map  $x$  to  $\hat{f}(x) \in \mathbb{R}^{k-1}$ .
- $\mathcal{A}$  is the class spaces as  $\mathcal{A} = \{1, 2, \dots, k\}$ , and  $a_i \in \mathcal{A}$  which is the class membership of subject  $i$ .
- We predict  $\hat{a}$  to be the class whose corresponding angle is the smallest, i.e.  $\hat{a} = \arg \min_j \angle(W_j, \hat{f})$ , where  $\angle(\cdot, \cdot)$  denotes the angle between two vectors.
- Minimizing the angle is equivalent to maximize  $\langle f(x_i), W_{a_i} \rangle$ .

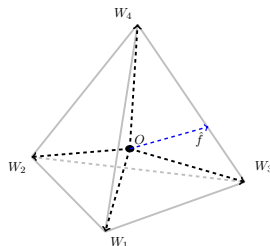
# Angle Based Classifier Illustration



(a)  $k = 2$



(b)  $k = 3$



(c)  $k = 4$

- $k = 2$ ,  $W_1 = 1$  and  $W_2 = -1$ .
- $k = 3$  (equilateral triangle),  
 $W_1 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$ ,  $W_2 = \left(\frac{\sqrt{3}-1}{2\sqrt{2}}, -\frac{\sqrt{3}+1}{2\sqrt{2}}\right)$ ,  $W_3 = \left(-\frac{\sqrt{3}+1}{2\sqrt{2}}, \frac{\sqrt{3}-1}{2\sqrt{2}}\right)$ .
- $k = 4$  (regular tetrahedron),  
 $W_1 = \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right)$ ,  $W_2 = \left(\frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}}\right)$ ,  $W_3 = \left(-\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}}\right)$ ,  $W_4 = -\left(\frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right)$ .

# Angle Based Classifier

With  $\ell$  a convex monotone decreasing function, we have our angle based classifier as,

$$\underset{f \in F}{\text{minimize}} \frac{1}{n} \sum_{i=1}^n \ell\{\langle f(x_i), W_{a_i} \rangle\} + \lambda J(f). \quad (5)$$

## Example ( $k = 2$ )

For a binary case, i.e.  $k = 2$ ,  $\langle f(x_i), W_{a_i} \rangle = af(x_i)$ ,

- When  $\ell(\cdot)$  is a deviance loss,  $\ell(z) = \log\{1 + \exp(-z)\}$ , equation (5) is a logistic regression.
- When  $\ell(\cdot)$  is a hinge loss,  $\ell(z) = (1 - z)_+$ , equation (5) is the support vector machine.

Original objective function,

$$D_o = \operatorname{argmin}_{D \in \mathcal{R}} n^{-1} \sum_{i=1}^n \frac{Y_i}{p(A_i|X_i)} I\{A_i \neq \mathcal{D}(X_i)\}.$$

ITR.ABC objective function (Zhang et al., 2017),

$$\underset{f \in F}{\text{minimize}} \frac{1}{n} \sum_{i=1}^n \frac{Y_i}{\Pr(A_i|X_i)} \ell\{\langle f(x_i), W_{a_i} \rangle\} + \lambda J(f).$$

Ordinal Extension (multiple treatments have a natural order): Chen et al. (2017)