Inferring causality from passive observations

Dominik Janzing

Max Planck Institute for Intelligent Systems
Tübingen, Germany

18.-22. August 2014

MAX-PLANCK-GESELLSCHAFT

Outline

1. **why the relation between statistics and causality is tricky**
2. **causal inference using conditional independences (statistical and general)**
3. **causal inference using other properties of joint distributions**
4. **causal inference in time series, quantifying causal strength**
5. **why causal problems matter for prediction**

Part 3: Why causality matters for prediction

- **Typical scenarios of machine learning**
  supervised learning, unsupervised learning, semi-supervised learning
- **Causality and semi-supervised learning**
- **Some methods of machine learning**
  support vector machines, reproducing kernel Hilbert spaces
- **Prediction from many variables**
  Markov blanket
- **Kind of summary**

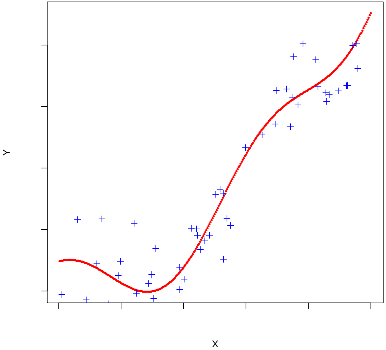
Typical scenarios of machine learning

Typical machine learning problems

- read hand-written letters/digits (given as pixel vectors)
- classify the topics of documents (i.e. in the internet)
- predict the structure of proteins from their acid sequence
- classify cell tissue into tumors and healthy tissue from gene expression data

Scenario with i.i.d. data

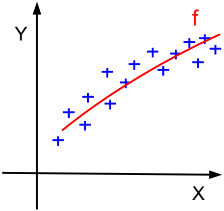
Task: predict y from x after observing some samples i.i.d. drawn from $P(X, Y)$.



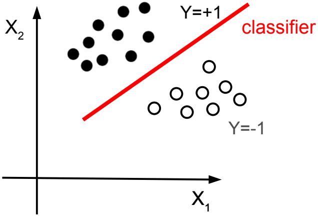
The variables may attain values in a high-dimensional space (e.g. pixel vector in $\mathbb{R}^{20 \cdot 20}$, gene expression vector in $\mathbb{R}^{28,000}$)

Regression and classification

- Regression: continuous label Y



- Classification: discrete label Y



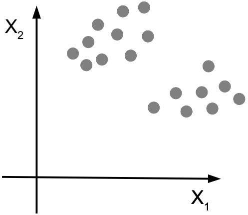
Supervised, unsupervised and semi-supervised learning

e.g. Chapelle et al: Semi-supervised learning, MIT press, 2006

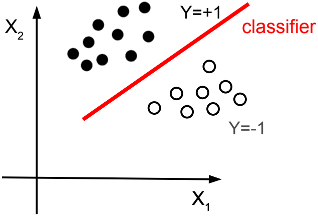
Learn to predict Y from X, given some observations. Which observations are given?

- **supervised learning:**
  some pairs $(x_1, y_1), \ldots, (x_n, y_n)$
- **unsupervised learning:**
  only unlabelled x-values $x_1, \ldots, x_k$
- **semi-supervised learning (SSL):**
  some pairs $(x_1, y_1), \ldots, (x_n, y_n)$ and some unlabelled x-values $x_{n+1}, \ldots, x_{n+k}$

Unsupervised classification

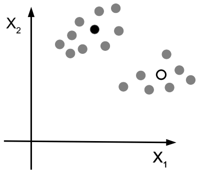


separate clusters are believed to correspond to different labels y:

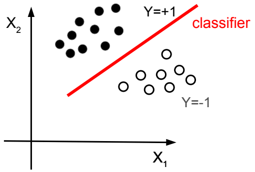


Semi-supervised classification

some points are labeled:

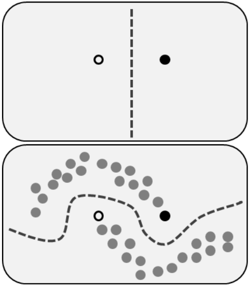


then we believe even more that the clusters correspond to different labels:



unlabeled points tell us position and angle of the hyperplane

Another example of semi-supervised classification...

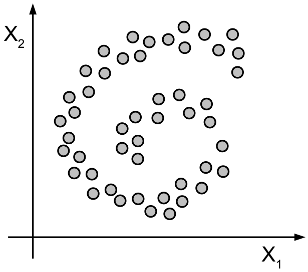


source: http://en.wikipedia.org/wiki/Semi-supervised_learning

Here, the shape of the clusters tells us something about the shape of the classifier.

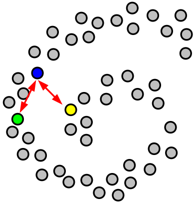
Semi-supervised regression

- **Given:** multi-dimensional $\mathbf{X}$ and real-valued Y
- **Task:** infer function f such that $y \approx f(\mathbf{x})$
- **Observation:** x-values essentially form a lower dimensional manifold of a higher dimensional space



Semi-supervised smoothness assumption

- Assumption: the map $\mathbf{x} \mapsto f(\mathbf{x})$ changes smoothly *along* the manifold



- Hence: the green and the yellow pojnts have the same distance to the blue one, the green one is more likely to have a similar y
- semi-supervised regression tries to find functions that fit the labeled data while changing smoothly *along the manifold*

Motivation for semi-supervised learning

- often it's expensive to get labeled data
- while unlabeled ones are cheap

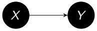
Examples:

- classify text w.r.t. its topic: labeling needs to be done by humans. Unlabeled documents are ubiquitous (e.g. in the internet)
- object recognition on images

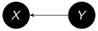
Causality and semi-supervised learning

SSL in Causal and Anti-Causal settings

The task is to predict Y from X



causal setting: predict effect from cause
e.g., predict splice sites from DNA sequence



anticausal setting: predict cause from effect
e.g, breast tumor classification,
image segmentation

Schölkopf, Janzing, Peters, Sgouritsa, Zhang, Mooij: On causal and anticausal learning, ICML 2012

SSL in Causal and Anti-Causal settings

The task is to predict Y from X

$X \rightarrow Y$

causal setting: predict effect from cause
e.g., predict splice sites from DNA sequence

$X \leftarrow Y$

anticausal setting: predict cause from effect
e.g, breast tumor classification,
image segmentation

SSL pointless because $P(X)$ contains no information about $P(Y|X)$

SSL can help because $P(X)$ and $P(Y|X)$ may contain information about each other

Schölkopf, Janzing, Peters, Sgouritsa, Zhang, Mooij: On causal and anticausal learning, ICML 2012

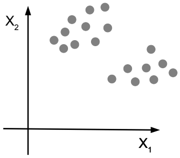
Known SSL assumptions link $P(X)$ to $P(Y|X)$

- *SSL smoothness assumption*: the function $x \mapsto \mathbb{E}[Y|x]$ should be smooth in regions where $p(x)$ is large.
- *Cluster assumption*: points lying in the same cluster are likely to have the same Y.
- *Low density separation*: The decision boundary should lie in a region where $p(x)$ is small.

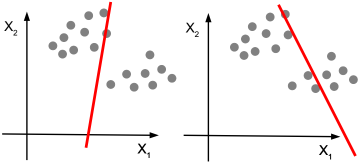
The above assumptions can indeed be viewed as **linking properties of $P(X)$ to properties of $P(Y|X)$**.

Consider classification in causal direction

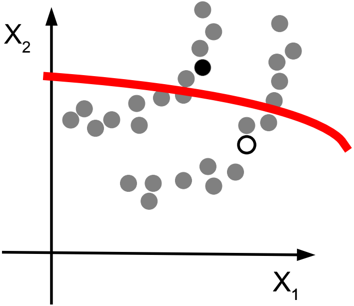
- Assume X is the cause with some distribution $P(X)$:



- Let Y be binary and $P(Y|X)$ be some deterministic rule, e.g., a separating hyperplane. Assume 'nature chooses' it independently of $P(X)$:



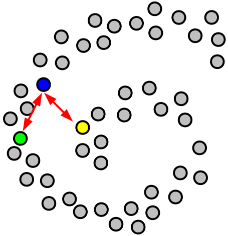
Consider curved clusters...



why should the shape of the line describing $P(Y|X)$ care about the shape of $P(X)$?

Recall semi-supervised regression

We said: close points should have similar values $f(x)$ if they have a short connection *along the manifold*



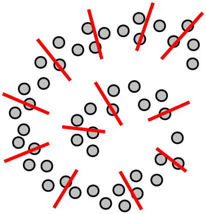
Can be justified if $Y \to \mathbf{X}$

Let Y and $\mathbf{X}$ take values in $\mathbb{R}$ and $\mathbb{R}^2$, respectively and

$$\mathbf{X} = (g_1(Y), g_2(Y)) + \mathbf{U}\,,$$

where g_1, g_2 are continuous functions and $\mathbf{U}$ is a small noise vector.

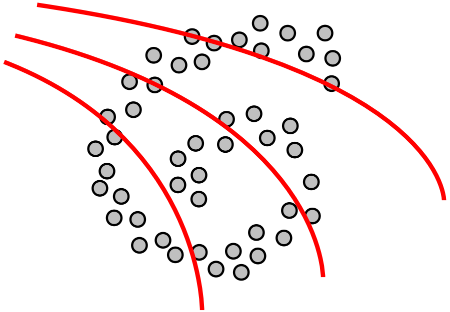
if y is close to y' then x, x' are close *along the manifold*



(isolines could be like the red lines)

Not justified for $\mathbf{X} \rightarrow Y$

Choose a function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ independent of $P(X)$



Then the isolines of f are nor related to the manifold on which the x-values lie

Does SSL only work in anticausal direction?

Meta-study:

- we searched the literature for empirical results on SSL with real data
- we decided (after several discussions) whether a data pair is causal or anticausal or unclear
  (here strongly confounded pairs were also considered anticausal)
- it turned out that none of the sucessful cases was causal

Schölkopf, Janzing, Peters, Sgouritsa, Zhang, Mooij: On causal and anticausal learning, ICML 2012

Semi-supervised classification: 8 benchmark datasets

Table 1. Categorization of eight benchmark datasets of Section 5 (Semi-supervised classification) as Anticausal/Confounded, Causal or Unclear

Category	Dataset	Reason of categorization
Anticausal/ Confounded	g241c	The class causes the 241 features.
	g241d	The class (binary) and the features are confounded by a variable with 4 states.
	Digit1	The positive or negative angle and the features are confounded by the variable of continuous angle.
	USPS	The class and the features are confounded by the 10-state variable of all digits.
	COIL	The six-state class and the features are confounded by the 24-state variable of all objects.
Causal	SecStr	The amino acid is the cause of the secondary structure.
Unclear	BCI, Text	Unclear which is the cause and which the effect.

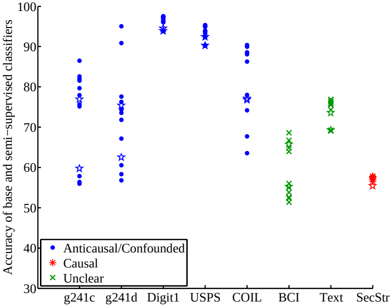
[O. Chapelle *et al.*, 2006] Semi-Supervised Learning.

Semi-supervised classification: 8 benchmark datasets

Table 1. Categorization of eight benchmark datasets of Section 5 (Semi-supervised classification) as Anticausal/Confounded, Causal or Unclear

Category	Dataset	Reason of categorization
Anticausal/ Confounded	g241c	The class causes the 241 features.
	g241d	The class (binary) and the features are confounded by a variable with 4 states.
	Digit1	The positive or negative angle and the features are confounded by the variable of continuous angle.
	USPS	The class and the features are confounded by the 10-state variable of all digits.
	COIL	The six-state class and the features are confounded by the 24-state variable of all objects.
Causal	SecStr	The amino acid is the cause of the secondary structure.
Unclear	BCI, Text	Unclear which is the cause and which the effect.

[O. Chapelle *et al.*, 2006] Semi-Supervised Learning.



Comparison of 11 SSL methods with the base classifiers 1-NN and SVM (star).

- In the causal case (red) the star is below the other points, i.e., SSL performed worse than methods without SSL
- In the other cases the star is often above the other points, SSL helped.

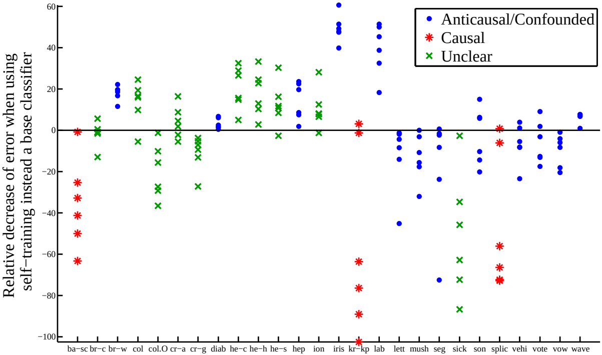
Semi-supervised classification: 26 UCI datasets

Table 2. Categorization of 26 UCI datasets of Section 5 (Semi-supervised classification) as Anticausal/Confounded, Causal or Unclear

Category	Dataset	Reason of categorization
Anticausal/ Confounded	breast-w	The class of the tumor (benign or malignant) causes some of the features of the tumor (e.g., thickness, size, shape etc.).
	diabetes	Whether or not a person has diabetes affects some of the features (e.g., glucose concentration, blood pressure), but also is an effect of some others (e.g. age, number of times pregnant).
	hepatitis	The class (die or survive) and many of the features (e.g., fatigue, anorexia, liver big) are confounded by the presence or absence of hepatitis. Some of the features, however, may also cause death.
	iris	The size of the plant is an effect of the category it belongs to.
	labor	Cyclic causal relationships: good or bad labor relations can cause or be caused by many features (e.g., wage increase, number of working hours per week, number of paid vacation days, employer's help during employee 's long term disability). Moreover, the features and the class may be confounded by elements of the character of the employer and the employee (e.g., ability to cooperate).
	letter	The class (letter) is a cause of the produced image of the letter.
	mushroom	The attributes of the mushroom (shape, size) and the class (edible or poisonous) are confounded by the taxonomy of the mushroom (23 species).
	segment	The class of the image is the cause of the features of the image.
	sonar	The class (Mine or Rock) causes the sonar signals.
	vehicle	The class of the vehicle causes the features of its silhouette.
	vote	This dataset may contain causal, anticausal, confounded and cyclic causal relations. E.g., having handicapped infants or being part of religious groups in school can cause one's vote, being democrat or republican can causally influence whether one supports Nicaraguan contras, immigration may have a cyclic causal relation with the class. Crime and the class may be confounded, e.g., by the environment in which one grew up.
	vowel	The class (vowel) causes the features.
	waveform-5000	The class of the wave causes its attributes.
Causal	balance-scale	The features (weight and distance) cause the class.
	kr-vs-kp	The board-description causally influences whether white will win.
	splice	The DNA sequence causes the splice sites.
Unclear	breast-cancer, colic, colic.ORIG, credit-a, credit-g, heart-c, heart-h, heart-statlog, ionosphere, sick	In some of these datasets, it is unclear whether the class label has been generated or defined based on the features (e.g., Ionoshpere, Credit Approval, Sick).

[Y. Guo *et al.*, 2006] An extensive empirical study on semi-supervised learning.

Semi-supervised classification: 26 UCI datasets



Comparison of a SSL method with 6 corresponding non-SSL-based classifiers.

- all red points are below or on the zero line, i.e., SSL performed worse or equal than all the non-SSL-based methods
- blue and green points are sometime above the line, i.e., SSL helped.

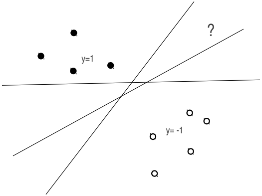
to understand *how* SSL works...

Some basic ideas of machine learning

Binary classification

Examples:

- distinguish between 0 and 1 from given pixel vectors in hand written digits
- distinguish between healthy cells and tumor cells given the vector of gene expression levels



- **given:** some training examples $(x_1, y_1), \ldots, (x_n, y_n)$ with $x_j \in \mathbb{R}^k$ and $y_j \in \{-1, +1\}$ (two classes)
- **goal:** find a function f that separates the two classes
- **assumption:** there is a linear classifier, i.e., an affine f such that $f(x_j) > 0$ if and only $y_j = +1$

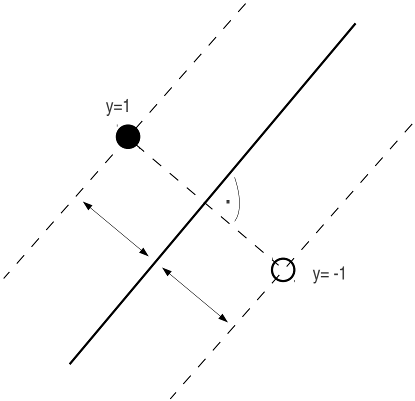
Popular solution: support vector machine

Vapnik 1963

see also Schölkopf & Smola: Learning with kernels

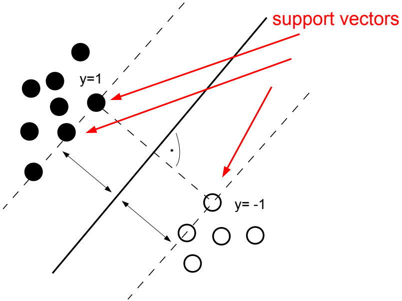
choose the hyperplane that maximizes the margin:
find $\mathbf{w} \in \mathbb{R}^k$ and $b \in \mathbb{R}$ that minimize $\|\mathbf{w}\|^2$ subject to

$$y_j(\langle \mathbf{w}, \mathbf{x}_j \rangle + b) \geq 1 \quad \forall j = 1, \ldots, n$$



Support vectors

The support vectors are those $\mathbf{x}_j$ that lie on the boundary of the margin:

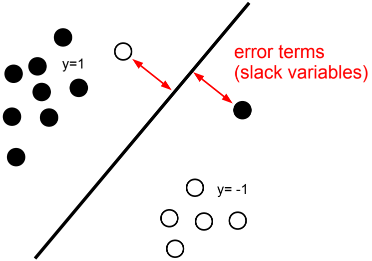


observations:

- points other than support vectors are irrelevant for the optimization
- the optimal $\mathbf{w}$ lies in the span of the support vectors

Soft margin SVMs

Often there is no hyperplane that classifies all points correctly:

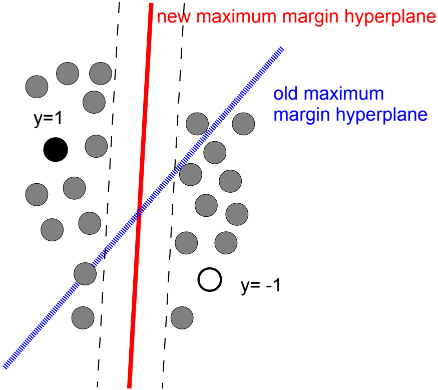


maximize margin while avoiding too large error terms

How can we include unlabelled points?

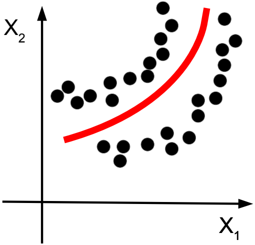
Transductive support vector machine

see e.g. Joachims: Transductive support vector machines, 2006



maximize the margin using also the unlabelled points
(yields computationally hard optimization problem)

Non-linear classification



here we need a non-linear function f for classification

In other words: find a function f for which $f(\mathbf{x}_j) > 0$ if $y_j = +1$ and $f(\mathbf{x}_j) < 0$ if $y_j = -1$.

Non-linear classification via feature maps

- define a map ϕ that maps x to non-linear functions of x:

$$\phi(x) := \begin{pmatrix} f_1(x) \\ f_2(x) \\ f_3(x) \\ . \\ . \\ . \end{pmatrix} \in \mathcal{H}$$

- in this high-dimensional space $\mathcal{H}$, a non-linear function may turn into a linear one
- example: let $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$ and define

$$\phi(\mathbf{x}) := \begin{pmatrix} x_1^2 \\ x_2^2 \end{pmatrix} .$$

Then the circle $x_1^2 + x_2^2 = 1$ is given by the linear equation

$$\phi_1(\mathbf{x}) + \phi_2(\mathbf{x}) = 1$$

‘Powerful’ feature maps

- we don’t need to design a different feature map for every problem
  (for instance one that already knows the classifier)
- it will turn out that there are feature maps that are so powerful that they can always be used
- idea: choose a map that contains an infinite set of functions, e.g.,

$$\phi(x) = \begin{pmatrix} x \\ x^2 \\ x^3 \\ \cdot \\ \cdot \\ \cdot \end{pmatrix} .$$

The kernel trick: we don't need ϕ itself...

Define the function

$$k(\mathbf{x}, \mathbf{x}') := \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle \,.$$

- recall optimization problem:
  minimize $\|\mathbf{w}\|^2$ subject to

$$y_j(\langle \mathbf{w}, \mathbf{x}_j \rangle + b) \geq 1 \quad \forall j = 1, \ldots, n$$

- recall that the optimal $\mathbf{w}$ can be written as

$$\mathbf{w} = \sum_j a_j \phi(\mathbf{x}_j)$$

  with some appropriate a
- to classify a new point $\mathbf{x}$ we only need to compute

$$\sum_j a_j \langle \phi(\mathbf{x}_j), \phi(\mathbf{x}) \rangle = \sum_j a_j k(\mathbf{x}_j, \mathbf{x}) \,.$$

General remarks

- Many other machine learning algorithms also rely on inner products only
- Then it is sufficient to know the 'kernel' k instead of knowing ϕ explicitly
- k can be rather simple functions even for complex ϕ.
  Example: for

$$\phi(x) := \begin{pmatrix} 1 \\ \sqrt{\binom{d}{1}}x \\ \sqrt{\binom{d}{2}}x^2 \\ \cdot \\ \cdot \\ \sqrt{\binom{d}{d}}x^d \end{pmatrix},$$

  we have

$$k(x, x') = \langle \phi(x), \phi(x') \rangle = (xx' + 1)^d \,.$$

Define vector spaces by kernels

Turning around the idea: define a vector space via the kernel.

Question:
Given a function

$$(x, x') \mapsto k(x, x'),$$

under which condsitions is there a map ϕ such that

$$k(x, x') = \langle \phi(x), \phi(x') \rangle \,?$$

Positive definite functions

Definition

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a positive definite kernel if for all $k \in \mathbb{N}$ and any $x_1, \ldots, x_k$, the matrix $k(x_i, x_j)$ is positive definite, i.e.,

$$\sum_{ij} c_i c_j k(x_i, x_j) \geq 0 \,,$$

for all vectors $c \in \mathbb{R}^k$.

Idea: positive definite functions are like positive definite matrices with continuous index x

Feature maps define positive definite kernels

Theorem

Let $\mathcal{H}$ be a real-valued vector space with inner product and and $\phi : \mathcal{X} \to \mathcal{H}$ be a map then

$$k(x_i, x_j) := \langle \phi(x_i), \phi(x_j) \rangle$$

is a positive definite kernel

Proof:

$$\begin{aligned} & \sum_{ij} c_i c_j \langle \phi(x_i), \phi(x_j) \rangle \\ = \; & \left\langle \sum_i c_i \phi(x_i), \sum_j c_j \phi(x_j) \right\rangle \geq 0 \, . \end{aligned}$$

Terminology: Hilbert spaces

- (possibly infinite-dimensional) vector space $\mathcal{H}$
- endowed with an inner product

$$\langle ., . \rangle \quad : \quad \mathcal{H} \times \mathcal{H} \to \mathbb{R}$$
$$\langle ., . \rangle \quad : \quad \mathcal{H} \times \mathcal{H} \to \mathbb{C}$$

- complete w.r.t. inner product norm, i.e., every Cauchy sequence in $\mathcal{H}$ has a limit in $\mathcal{H}$

Examples: $\mathbb{C}^n, \mathbb{R}^n, L^2(\mathbb{R}), \ell^2(\mathbb{N})$

Defining Hilbert spaces via kernels

Theorem (Proposition 2.14 in Schölkopf & Smola, 2002)

Let $\mathcal{X}$ be a topological space and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous positive definite kernel, then there exists a Hilbert space $\mathcal{H}$ and a continuous map $\phi : \mathcal{X} \to \mathcal{H}$ such that for all $x, x' \in \mathcal{X}$, we have

$$k(x, x') = \langle \phi(x), \phi(x') \rangle .$$

Hence: feature maps ϕ define kernels and vice versa

Examples for kernels

- **Gauss kernel:** $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$

$$k(\mathbf{x}, \mathbf{x}') := e^{\frac{-\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}} .$$

- **polynomial kernel:**

$$k(\mathbf{x}, \mathbf{x}') := \langle \mathbf{x}, \mathbf{x}' \rangle^d$$

- **inhomogeneous polynomial kernel:**

$$k(\mathbf{x}, \mathbf{x}') := (\langle \mathbf{x}, \mathbf{x}' \rangle + 1)^d$$

Reproducing kernel Hilbert space (RKHS)

How to get $\mathcal{H}$ from k:

- Every x defines a function $f_x : \mathcal{X} \mapsto \mathbb{R}$ via

$$f_x(x') := k(x, x')\,.$$

- Define an inner product by

$$\langle f_x, f_{x'} \rangle := k(x, x')\,.$$

- Let $\mathcal{H}$ be the smallest Hilbert space containing all these functions $(f_x)_{x \in \mathcal{X}}$ with the above inner product

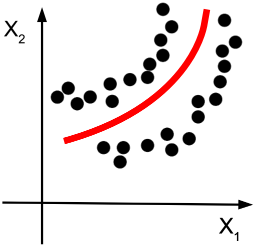
Gauss kernel is powerful

Define

$$k(\mathbf{x}, \mathbf{x}') := e^{\frac{-\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}},$$

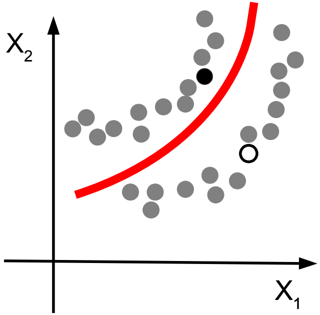
then there is a sense in which every continuous function can be approximated by elements in $\mathcal{H}$

Therefore, support vector machine with the Gaussian kernels are able to find curved separating hyperplanes:



Likewise...

Transductive SVM with Gauss kernels is able to find curved separating hyperplanes based on the pattern of the unlabeled points:



Applications of SSL

e.g. Nigam et al: Semi-supervised text classification using EM, 2006

Text classification

- **task:** assign a topic y to a document x in the internet
- **given:** a few labelled documents $(x_1, y_1), \dots, (x_n, y_n)$ (expensive, done by humans), and a huge number of unlabelled examples $x_{n+1}, \dots, x_{n+k}$
- **idea:** different topics correspond to clusters in high-dimensional space
  (x vector of word counts, i.e., $x \in \mathbb{N}_0^d$ where d is the size of the vocabulary)

Prediction from many variables

Prediction tasks where causal information may be irrelevant

Health Insurance company asks a costumer before giving a contract:

- age
- job
- hobbies
- address

to determine the risk of getting sick.

It does not matter whether these features are causal or not - properties that correlate with the risk are useful regardless of they are causes!

Markov blanket (set of relevant variables)

- **given:** causal DAG with nodes $X_1, \ldots, X_n$
- **task:** predict X_j from all the other variables
- **obvious solution:** best prediction given by

$$p(X_j | X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n) \,.$$

- **goal:** remove irrelevant nodes
- **question:** what's the smallest set of variables S such that

$$X_j \perp\!\!\!\perp \mathbf{X}_S \,| \mathbf{X}_{\bar{S}} \setminus X_j \,?$$

- **answer:** Markov blanket, i.e., the parents and the children of X_j, and the other parents of the latter (if we assume faithfulness)

Proof that Markov blanket is the smallest set

- Let $M(j)$ be the Markov blanket of X_j. Then

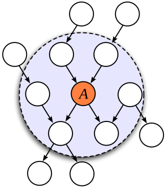
$$X_j \perp\!\!\!\perp \mathbf{X}_{M(j)} \,|\mathbf{X}_{\bar{M(j)}} \setminus X_j \,, \tag{1}$$

  can be checked via d-separation

- every set S satisfying (1) contains $M(j)$:
  - no set can d-separate X_j from its parents or children. Therefore they are contained in S
  - conditioning on the children of X_j unblocks the path to the parents of the children, no matter which variables we condition on

Markov blanket Pearl 1988

- **Definition via causal terminology:**
  consists of parents, children and further parents of the children



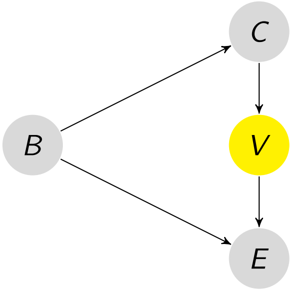
- **Definition via predictive relevance:**
  minimal set of variables that render all others irrelevant

Two perspectives:

- Causal structure tells us which variables are irrelevant because they are conditionally independent
- Conditional independences tell us causal structure

Argument for causality seems circular!

Predicting from causes is more robust



$P(V|C)$ predicts V from its cause ('causal prediction')

$P(V|E)$ predicts V from its effect ('anticausal prediction')

- $P(V|E)$ changes when background condition B changes because $V \not\perp\!\!\!\perp B \,|\, E$
- $P(V|C)$ remains constant ('covariate shift') since $V \perp\!\!\!\perp B \,|\, C$.

Schölkopf, Janzing, Peters, Sgouritsa, Zhang, Mooij *ICML 2012*

Kind of summary

Two types of asymmetries between cause and effect

- **Independence based assumption:** $P(C)$ and $P(E|C)$ contain no information about each other because 'nature chooses them independently'
  - no SSL in causal direction
  - $P(C)$ ansd $P(E|C)$ are algorithmically independent
  - $p(C)$ is not particualrly large in regions where f has large slope (IGCI)

- **Occam's Razor assumption:** Decomposition of $P(C, E)$ into $P(C)P(E|C)$ tends to be simpler than decomposition into $P(E)P(C|E)$
  - $P(E|C)$ may be of some simple type, e.g. non-linear additive noise, while $P(C|E)$ isn't
  - some Bayesian method not mentioned
  - $K(P(C)) + K(P(E|C)) \leq K(P(E)) + K(P(C|E))$

Relating the two types via a toy model

- choose $P(C)$ randomly from some finite set $P_i(C)$ with $i = 1, \ldots, n$
- choose $P(E|C)$ randomly from some finite set $P_j(E|C)$ with $j = 1, \ldots, n$
- then $P_i(C)P_j(E|C)$ defines n^2 different distributions $P_{ij}(C, E)$
- generically, this yields n^2 different distributions $P_{ij}(E)$ and $P_{ij}(C|E)$
- hence, the backwards conditionals $P(E)$ and $P(C|E)$ run over a larger set

Future work

find appropriate ways to compare complexities of $P(X), P(Y|X)$ to $P(Y)P(X|Y)$

Psychology

Humans intuition about **anticausal conditionals** $P(C|E)$ can be pretty bad:

- Consider a random walk on $\mathbb{Z}$
- Let X_t be the position at time t
- Let $p(X_{t+1}|X_t) = 1/2$ for $|X_{t+1} - X_t| = 1$
- Try to infer $p(X_t|X_{t+1})$

Psychology

Humans intuition about **anticausal conditionals** $P(C|E)$ can be pretty bad:

- Consider a random walk on $\mathbb{Z}$
- Let X_t be the position at time t
- Let $p(X_{t+1}|X_t) = 1/2$ for $|X_{t+1} - X_t| = 1$
- Try to infer $p(X_t|X_{t+1})$

- Many people think it would also be symmetric, i.e., given X_{t+1}, the two possibilities $X_t = X_{t+1} \pm 1$ would be equally likely

We are used to think in terms of causal conditionals, because they define the mechanisms

State of the art and outlook

- Asymmetries of cause and effect do exist (at least in the sense that decision rate above chance level is possible)
- Initial scepticism against the field has decreased
- Try your own ideas, good methods need not build upon the existing ones

Thank you for your attention