From Points to Measures
A Kernel Perspective

Krikamol Muandet Bernhard Schölkopf

MAX-PLANCK-GESELLSCHAFT

Max Planck Institute for
Intelligent Systems

Tübingen, Germany

BIOLOGISCHE KYBERNETIK

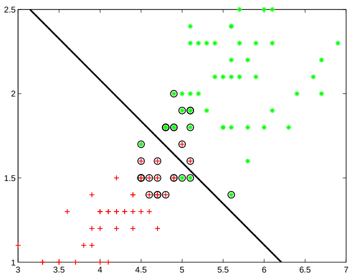
1. Learning from Data Points
2. Learning from Dirac Measures
3. Learning from Gaussian Measures

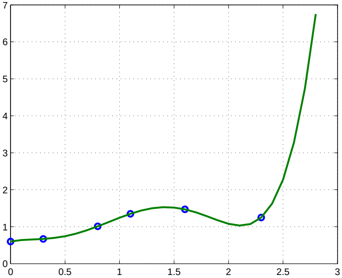
Learning from data points

Given finite samples $\{(x_i, y_i)\}_{i=1}^m$ drawn i.i.d. from $\mathcal{X} \times \mathcal{Y}$ according to $P(X, Y)$, the goal is to learn $f : \mathcal{X} \to \mathcal{Y}$ that encodes dependency between X and Y.

Learning from data points

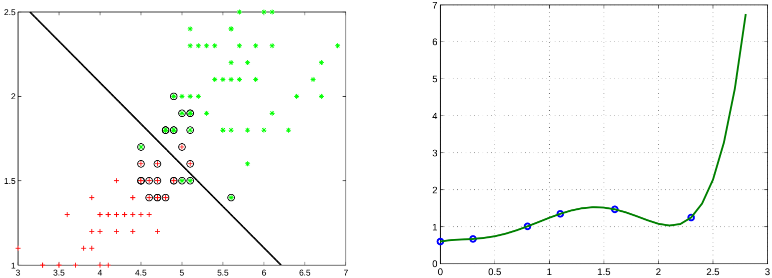
Given finite samples $\{(x_i, y_i)\}_{i=1}^m$ drawn i.i.d. from $\mathcal{X} \times \mathcal{Y}$ according to $P(X, Y)$, the goal is to learn $f : \mathcal{X} \to \mathcal{Y}$ that encodes dependency between X and Y.





Learning from data points

Given finite samples $\{(x_i, y_i)\}_{i=1}^m$ drawn i.i.d. from $\mathcal{X} \times \mathcal{Y}$ according to $P(X, Y)$, the goal is to learn $f : \mathcal{X} \rightarrow \mathcal{Y}$ that encodes dependency between X and Y.



Unfortunately, the dependency between X and Y is often *nonlinear*.

Learning from data points

The *kernel* method resolves this problem by considering a mapping

$$\phi : \mathcal{X} \to \mathcal{H}, \;\; x \longmapsto k(x, \cdot) \;\; ,$$

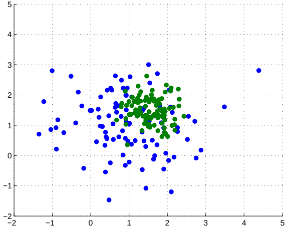
which embeds in some high-dimensional space $\mathcal{H}$ the set of data points.

Learning from data points

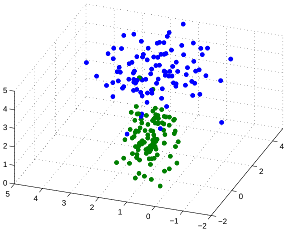
The *kernel* method resolves this problem by considering a mapping

$$\phi : \mathcal{X} \to \mathcal{H}, \ \ x \longmapsto k(x, \cdot) \ \ ,$$

which embeds in some high-dimensional space $\mathcal{H}$ the set of data points.



$\Longrightarrow$



Learning from data points

Theorem

Following the framework of Tikhonov regularization, any function $f \in \mathcal{H}$ minimizing the regularized risk functional

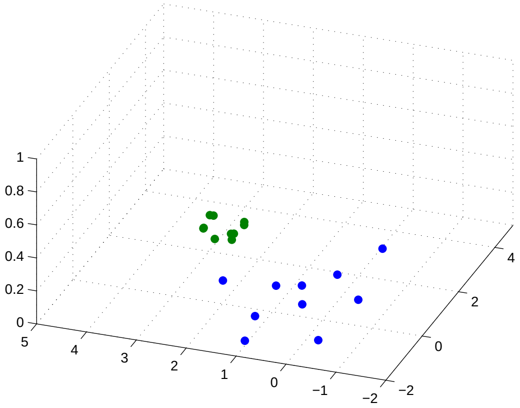
$$\mathcal{L}(\{x_i, y_i, f(x_i)\}_{i=1}^m) + \lambda\Omega(\|f\|_{\mathcal{H}})$$

admits the representation of the form

$$f = \sum_{i=1}^{m} \alpha_i k(x_i, \cdot)$$

for some $\boldsymbol{\alpha} \in \mathbb{R}^m$ and reproducing kernel k of $\mathcal{H}$.

Scenario 1 : Learning from Data Points



$$x \longmapsto k(x, \cdot)$$

1. Learning from Data Points
2. Learning from Dirac Measures
3. Learning from Gaussian Measures

Learning from dirac measures

Consider the Dirac measure δ_x on a measurable space $(\mathcal{X}, \mathcal{A})$, where $\mathcal{A}$ is a σ-algebra of subsets of $\mathcal{X}$, defined for x in $\mathcal{X}$ by

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

where $A \in \mathcal{A}$.

Learning from dirac measures

Consider the Dirac measure δ_x on a measurable space $(\mathcal{X}, \mathcal{A})$, where $\mathcal{A}$ is a σ-algebra of subsets of $\mathcal{X}$, defined for x in $\mathcal{X}$ by

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

where $A \in \mathcal{A}$. For any measurable function f on $\mathcal{X}$, we have

$$f(x) = \int f(t) \; d\delta_x(t)$$

Learning from dirac measures

Consider the Dirac measure δ_x on a measurable space $(\mathcal{X}, \mathcal{A})$, where $\mathcal{A}$ is a σ-algebra of subsets of $\mathcal{X}$, defined for x in $\mathcal{X}$ by

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

where $A \in \mathcal{A}$. For any measurable function f on $\mathcal{X}$, we have

$$f(x) = \int f(t) \; d\delta_x(t)$$

That is, the evaluation of f at point x is the expectation of f with respect to δ_x.

Learning from dirac measures

If $f \in \mathcal{H}$ of functions on $\mathcal{X}$ with reproducing kernel k, then

$$\langle f, k(x, \cdot)\rangle = \int f(t) \; d\delta_x(t) \; .$$

Learning from dirac measures

If $f \in \mathcal{H}$ of functions on $\mathcal{X}$ with reproducing kernel k, then

$$\langle f, k(x, \cdot) \rangle = \int f(t) \; d\delta_x(t) \; .$$

This defines a mapping

$$\phi : \mathcal{P} \to \mathcal{H}, \;\; \delta_x \longmapsto \mathbb{E}_{\delta_x}[k(x, \cdot)] \; ,$$

which embeds in $\mathcal{H}$ the set of Dirac measures on $\mathcal{X}$. It is trivial to see that this scenario is equivalent to **Scenario 1**.

Learning from dirac measures

If $f \in \mathcal{H}$ of functions on $\mathcal{X}$ with reproducing kernel k, then

$$\langle f, k(x, \cdot) \rangle = \int f(t) \ d\delta_x(t) \ .$$

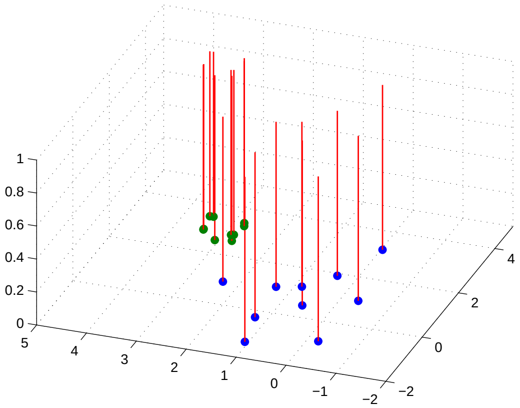
This defines a mapping

$$\phi : \mathcal{P} \to \mathcal{H}, \ \ \delta_x \longmapsto \mathbb{E}_{\delta_x}[k(x, \cdot)] \ ,$$

which embeds in $\mathcal{H}$ the set of Dirac measures on $\mathcal{X}$. It is trivial to see that this scenario is equivalent to **Scenario 1**.

This is in fact the motivation to embed the distributions into RKHS (Berlinet and Thomas-agnan, 2004; Smola et al., 2007).

Scenario 2 : Learning from Dirac Measures



$$\delta_x \longmapsto \mathbb{E}_{\delta_x}[k(x,\cdot)]$$

Scenario 1 $\equiv$ Scenario 2

Learning from dirac measures

Proposition

Let $\mathcal{F}$ be a set of functions in the reproducing kernel Hilbert space $\mathcal{H}$ having the form $f = \sum_{i=1}^{m} \alpha_i k(x_i, \cdot)$, where k is the reproducing kernel of $\mathcal{H}$, and $\mathcal{M}$ be a set of discrete signed measure $\mu = \sum_{i=1}^{m} \alpha_i \delta_{x_i}$ in $\mathcal{H}$. Then, for $m \geq 1$, we have

$$\mathcal{F} \equiv \mathcal{M} \quad .$$

Learning from dirac measures

Proposition

Let $\mathcal{F}$ be a set of functions in the reproducing kernel Hilbert space $\mathcal{H}$ having the form $f = \sum_{i=1}^{m} \alpha_i k(x_i, \cdot)$, where k is the reproducing kernel of $\mathcal{H}$, and $\mathcal{M}$ be a set of discrete signed measure $\mu = \sum_{i=1}^{m} \alpha_i \delta_{x_i}$ in $\mathcal{H}$. Then, for $m \geq 1$, we have

$$\mathcal{F} \equiv \mathcal{M} \quad .$$

In other words,

$$\sum_{i=1}^{m} \alpha_i k(x_i, \cdot) \equiv \sum_{i=1}^{m} \alpha_i \delta_{x_i}$$

Learning from dirac measures

Proof.

Any Hilbert space $\mathcal{H}$ of functions on $\mathcal{X}$ with reproducing kernel k contains, as a dense subset, the set $\mathcal{F}$ of linear combinations

$$\sum_{i=1}^{m} \alpha_i k(x_i, \cdot), \;\; m \geq 1, \;\; \alpha_i \in \mathbb{R}, \;\; x_i \in \mathcal{X},$$

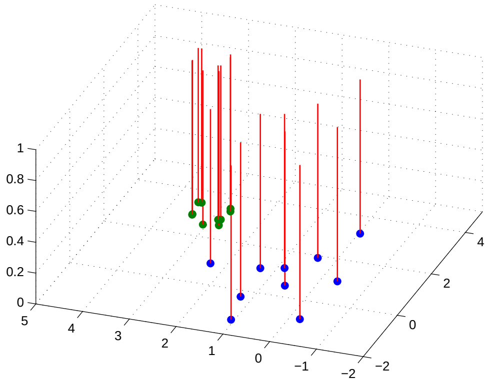
with the property that, for any measurable f in $\mathcal{H}$,

$$\langle f, \sum_{i=1}^{m} \alpha_i k(x_i, \cdot) \rangle = \sum_{i=1}^{m} \alpha_i f(x_i) = \int f \;\; d\mu$$

where $\mu = \sum_{i=1}^{m} \alpha_i \delta_{x_i}$ is the discrete signed measure putting the mass α_i at the point x_i.

□

Scenario 2 : Learning from Dirac Measures



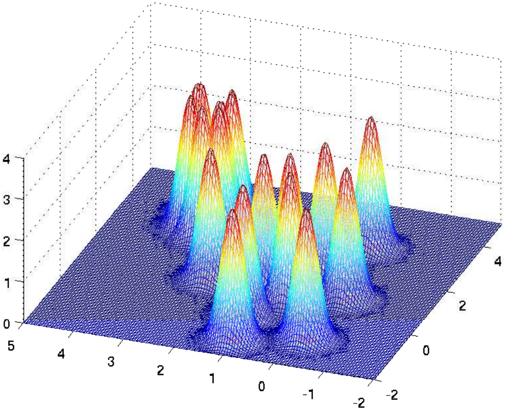
Regularization ≡ Finding the optimal linear combinations of Dirac measures $\{\delta_{x_1}, \delta_{x_2}, \dots, \delta_{x_m}\}$

1. Learning from Data Points

2. Learning from Dirac Measures

3. Learning from Gaussian Measures

Scenario 3 : Learning from Gaussian Measures



$$P \mapsto \mathbb{E}_P[k(x, \cdot)]$$

Learning from Gaussian Measures

Let $\mathcal{P}$ be a set of Gaussian probability measures P_σ with width σ and $\mathcal{H}_\sigma$ be a RKHS with Gaussian reproducing kernel k_σ. Define a map from $\mathcal{P}$ into $\mathcal{H}_\sigma$

$$\phi : \ \mathcal{P} \to \mathcal{H}_\sigma, \ \ P_\sigma \mapsto \mathbb{E}_{P_\sigma}[k_\sigma(x, \cdot)] \triangleq \mu[P_\sigma]$$

Learning from Gaussian Measures

Let $\mathcal{P}$ be a set of Gaussian probability measures P_σ with width σ and $\mathcal{H}_\sigma$ be a RKHS with Gaussian reproducing kernel k_σ. Define a map from $\mathcal{P}$ into $\mathcal{H}_\sigma$

$$\phi: \ \mathcal{P} \rightarrow \mathcal{H}_\sigma, \ \ P_\sigma \mapsto \mathbb{E}_{P_\sigma}[k_\sigma(x,\cdot)] \triangleq \mu[P_\sigma]$$

Due to the reproducing property of $\mathcal{H}_\sigma$, we have

$$\langle f, \mathbb{E}_P[k_\sigma(x,\cdot)]\rangle = \mathbb{E}_P[f(x)]$$

Learning from Gaussian Measures

Let $\mathcal{P}$ be a set of Gaussian probability measures P_σ with width σ and $\mathcal{H}_\sigma$ be a RKHS with Gaussian reproducing kernel k_σ. Define a map from $\mathcal{P}$ into $\mathcal{H}_\sigma$

$$\phi : \ \mathcal{P} \to \mathcal{H}_\sigma, \ \ P_\sigma \mapsto \mathbb{E}_{P_\sigma}[k_\sigma(x, \cdot)] \triangleq \mu[P_\sigma]$$

Due to the reproducing property of $\mathcal{H}_\sigma$, we have

$$\langle f, \mathbb{E}_P[k_\sigma(x, \cdot)] \rangle = \mathbb{E}_P[f(x)]$$

Then, define a set of functions

$$\mathcal{F} = \left\{ f \in \mathcal{H}_\sigma \middle| f(\cdot) = \sum_{i=1}^{\infty} \beta_i \mu[P_i], \beta_i \in \mathbb{R}, P_i \in \mathcal{P}, \|f\| < \infty \right\}$$

Learning from Gaussian Measures

Theorem

Given a training set $\{(x_i, y_i)\}_{i=1}^m$ from $\mathcal{X} \times \mathbb{R}$, a set of Gaussian probability measure $\{P_{\sigma_i}\}_{i=1}^m$ with density $\{p_{\sigma_i}\}_{i=1}^m$, a strictly monotonically increasing real-valued function Ω on $[0, \infty)$, arbitrary loss function $\mathcal{L} : (\mathcal{X} \times \mathbb{R}^2) \to \mathbb{R} \cup \{\infty\}$, and nonnegative regularization parameter λ, then any $f \in \mathcal{F}$ minimizing the regularized risk functional

$$\mathcal{L}\left(\{P_i, y_i, \mathbb{E}_{P_{\sigma_i}}[f(x)]\}_{i=1}^m\right) + \lambda\Omega(\|f\|)$$

admits a representation of the form

$$f(\cdot) = \sum_{i=1}^{m} \alpha_i k_i(x_i, \cdot)$$

where for some $\boldsymbol{\alpha} \in \mathbb{R}^m$ and $k_i = k_\sigma \otimes p_{\sigma_i}$.

Learning from Gaussian Measures

Proof.

Consider a bounded linear operator L_{P_i} such that $L_{P_i} f = \mathbb{E}_{P_i}[f(x)]$. Then it follows from Wahba (1990) that each solution f minimizing

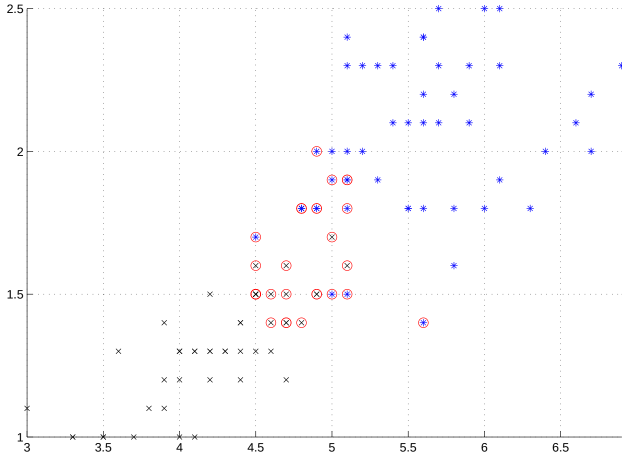
$$\mathcal{L}\left(\{P_i, y_i, \mathbb{E}_{P_{\sigma_i}}[f(x)]\}_{i=1}^m\right) + \lambda\Omega(\|f\|)$$

can be written as

$$f = \sum_{i=1}^m \alpha_i k_i(\cdot)$$

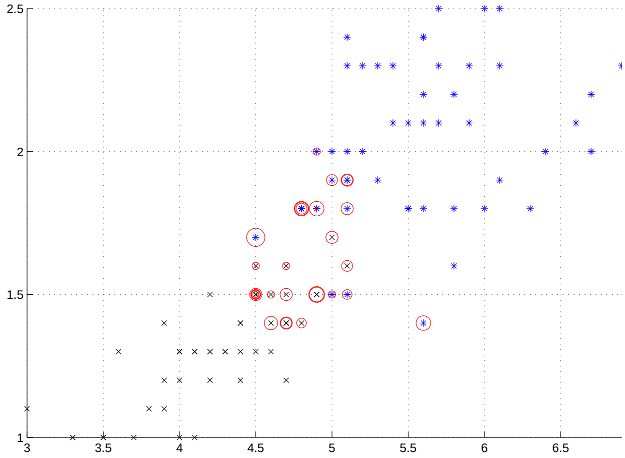
where each $k_i(\cdot)$ corresponds to each L_{P_i}. □

Application



SVM with fixed widths

Application



SVM with variable widths

Related Works

Related Works

- Probability Product Kernel (Jebara et al., 2004)

Related Works

- Probability Product Kernel (Jebara et al., 2004)
  - Fitting probabilistic models $p_1(x), \dots, p_m(x)$ to $x_1, \dots, x_m$.

Related Works

- Probability Product Kernel (Jebara et al., 2004)
  - Fitting probabilistic models $p_1(x), ..., p_m(x)$ to $x_1, ..., x_m$.
  - Define kernel $k^*(p, p')$ between probability distributions on $\mathcal{X}$.

Related Works

- Probability Product Kernel (Jebara et al., 2004)
  - Fitting probabilistic models $p_1(x), ..., p_m(x)$ to $x_1, ..., x_m$.
  - Define kernel $k^*(p, p')$ between probability distributions on $\mathcal{X}$.
  - Define the kernel between examples to equal k^* between the corresponding distributions:

$$k(x, x') = k^*(p, p')$$

Related Works

- Probability Product Kernel (Jebara et al., 2004)
  - Fitting probabilistic models $p_1(x), \ldots, p_m(x)$ to $x_1, \ldots, x_m$.
  - Define kernel $k^*(p, p')$ between probability distributions on $\mathcal{X}$.
  - Define the kernel between examples to equal k^* between the corresponding distributions:

$$k(x, x') = k^*(p, p')$$

- Learning using Previleged Information (LUPI) (Pechyony and Vapnik, 2010; Vapnik and Vashist, 2009)

Related Works

- Probability Product Kernel (Jebara et al., 2004)
  - Fitting probabilistic models $p_1(x), \ldots, p_m(x)$ to $x_1, \ldots, x_m$.
  - Define kernel $k^*(p, p')$ between probability distributions on $\mathcal{X}$.
  - Define the kernel between examples to equal k^* between the corresponding distributions:

$$k(x, x') = k^*(p, p')$$

- Learning using Previleged Information (LUPI) (Pechyony and Vapnik, 2010; Vapnik and Vashist, 2009)
  - In addition to training data $\{(x_i, y_i)\}_{i=1}^m$, the privileged information $x^* \in \mathcal{X}^*$ is also available.

Related Works

- Probability Product Kernel (Jebara et al., 2004)
  - Fitting probabilistic models $p_1(x), \dots, p_m(x)$ to $x_1, \dots, x_m$.
  - Define kernel $k^*(p, p')$ between probability distributions on $\mathcal{X}$.
  - Define the kernel between examples to equal k^* between the corresponding distributions:

$$k(x, x') = k^*(p, p')$$

- Learning using Previleged Information (LUPI) (Pechyony and Vapnik, 2010; Vapnik and Vashist, 2009)
  - In addition to training data $\{(x_i, y_i)\}_{i=1}^m$, the privileged information $x^* \in \mathcal{X}^*$ is also available.
  - The privileged information is only available for the training examples.

Related Works

- Probability Product Kernel (Jebara et al., 2004)
  - Fitting probabilistic models $p_1(x), \dots, p_m(x)$ to $x_1, \dots, x_m$.
  - Define kernel $k^*(p, p')$ between probability distributions on $\mathcal{X}$.
  - Define the kernel between examples to equal k^* between the corresponding distributions:

$$k(x, x') = k^*(p, p')$$

- Learning using Previleged Information (LUPI) (Pechyony and Vapnik, 2010; Vapnik and Vashist, 2009)
  - In addition to training data $\{(x_i, y_i)\}_{i=1}^m$, the privileged information $x^* \in \mathcal{X}^*$ is also available.
  - The privileged information is only available for the training examples.
- Gaussian Processes

Summary

1. Learning from Data Points
2. Learning from Dirac Measures
3. Learning from Gaussian Measures

Acknowledgement

- Christian Walder
- Samory Kpotufe
- Francesco Dinuzzo

References

Berlinet, A. and C. Thomas-agnan (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers.

Jebara, Tony et al. (2004). "Probability product kernels". In: *Journal of Machine Learning Research* 5, pp. 819–844.

Pechyony, Dmitry and Vladimir Vapnik (2010). "On the Theory of Learning with Privileged Information". In: *Advances in Neural Information Processing Systems 23*.

Smola, Alex et al. (2007). "A Hilbert space embedding for distributions". In: *In Algorithmic Learning Theory: 18th International Conference*. Springer-Verlag, pp. 13–31.

Vapnik, Vladimir and Akshay Vashist (2009). "A new learning paradigm: Learning using privileged information". In: *Neural Networks* 22.5-6, pp. 544–557.

Wahba, G. (1990). *Spline models for Observational data (CBMS-NSF Regional Conference Series in Applied Mathematics)*. Philadelphia: Society for Industrial and Applied Mathematics, p. 180.

Questions & Comments?