1 Recap from last lecture

1.1 Convex optimisation principals

1.2 Non parametric estimation

1.2.1 Nearest neighbour algorithm (NN)

1.2.2 K-NN

1.2.3 Smoothing technique for regression

2 New stuff

2.1 Theorem Cover and Hart 1967

2.2 Nonlinear classification with kernels

2.2.1 Introduction

2.2.2 RKHS (Reproducing Kernel Hilbert Space) and kernels

1 Recap from last lecture

n is the number of training points, p the number of dimensions of the points and M the number of labels

1.1 Convex optimisation principals

- Gradient descent algorithm

\[
\min_{\theta \in \mathbb{R}^p} f(\theta)
\]

\[
\theta_{t+1} \leftarrow \theta_t - \eta \nabla f(\theta_t)
\]

- Newton algorithm

\[
\theta_{t+1} \leftarrow \theta_t - \nabla^2 f(\theta_t)^{-1} \nabla f(\theta_t)
\]
• Projected gradient descent

\[
\min_{\theta \in \mathcal{C}} f(\theta)
\]

\[
\theta_{t+1} \leftarrow \pi_{\mathcal{C}}[\theta_t - \eta_t \nabla f(\theta_t)]
\]

• Proximal gradient descent

\[
\min_{\theta \in \mathbb{R}^p} (f(\theta) + \Omega(\theta))
\]

\[
\theta_{t+1} \leftarrow \arg \min_{\theta \in \mathbb{R}^p} \frac{1}{2} \| \theta - \left[ \theta_t - \frac{1}{L} \nabla f(\theta_t) \right] \|^2_L + \frac{1}{L} \Omega(\theta)
\]

• Stochastic gradient descent

\[
\min_{\theta \in \mathbb{R}^p} E_x[l(\theta, x)]
\]

1.2 Non parametric estimation

1.2.1 Nearest neighbour algorithm (NN)

Training data are \((X_i, X_j)_{i=1,...,n}\), where \(X_i \in \mathbb{R}^p\) and \(X_j \in \{1, \ldots, M\}\)

Given a test point \(X \in \mathbb{R}^p\)

\[
\hat{i}(X) = \arg \min_{i=1,...,n} d(X, X_i)
\]

\[
\hat{y}_{NN}(X) = y_{\hat{i}(X)}
\]

1.2.2 K-NN

Extension of voting scheme for the K nearest neighbours

\[
\hat{y}_{k-NN}(X) = Vote(y_{i_1}, \ldots, y_{i_k})
\]

1.2.3 Smoothing technique for regression

Training data are \((X_i, Y_i)\), where \(X_i \in \mathbb{R}^p\) and \(Y_i \in \mathbb{R}\)

\[
\hat{y}(X) = \sum_{i=1}^n \frac{K_\sigma(X, X_i)Y_i}{\sum_{j=1}^n K_\sigma(X_j, X_i)}
\]
2 New stuff

2.1 Theorem Cover and Hart 1967

“asymptotically, the NN error rate is never more than twice the Bayes error rate” [2], whereby the Bayes error rate is the best achievable result.

The training data is \((Y_i, X_i)\) with \(Y_i\) drawn according to \(P[Y = c|X]\), where \(c \in \{1, \ldots, M\}\).

\[
\min_{f: \mathbb{R}^p \to \{1, \ldots, M\}} E_{(X,Y)}[\mathbb{1}_{f(X) \neq Y}] = \\
E_X E_{Y|X}[\mathbb{1}_{f(X) \neq Y}] = \\
E_X \left[ \sum_{c=1}^{n} P[Y = c|X] \mathbb{1}_{f(X) \neq Y} \right] = \\
E_X \left[ \sum_{c=1}^{n} P[Y = c|X] - P[Y = f(X)|X] \right] = \\
E_X [1 - P[Y = f(X)|X]]
\]

The Bayes classifier minimizes the above quantity. It is such that

\[
\hat{Y}_{\text{Bayes}}(X) = \arg \max_{c=1, \ldots, n} P[Y = c|X]
\]

**Sketch of proof:** (proven in the ’60s)

A) Prove theorem in some “ideal” setting

B) Show that the ideal case “converges” to the general case (very technical proof)

We only do A)

Ideal setting means two things:

- \(M = 2\) (two labels)
- Assume that the training set is infinite and dense
  \(\forall X \in \mathbb{R}^p\) there exists \((X_i, Y_i)\) in the training set with \(X = X_i\)
Error rate: \( E(Y, X, \text{training data})[1_{Y \neq \hat{Y}}(X)] \), where \((Y, X)\) is the test data

\[
\begin{align*}
E(Y, X, Y')[1_{Y \neq Y'}] & = E(Y, X) \left[ \sum_{c=1}^{M=2} P[Y' = c|X] \ 1_{Y \neq c} \right] \\
& = E_X \left[ \sum_{c=1}^{2} P[Y' = c|X] \ E_{Y'|X}[1_{Y \neq c}] \right] \\
& = E_X \left[ \sum_{c=1}^{2} P[Y = c|X] (1 - P[Y = c|X]) \right] \\
& = E_X \left[ 2P[Y \neq \hat{Y}_{\text{Bayes}}(X)|X] (1 - \frac{P[Y \neq \hat{Y}_{\text{Bayes}}(X)|X]}{\leq 1}) \right] \\
& \leq 2E_X \left[ P[Y \neq \hat{Y}_{\text{Bayes}}(X)|X] \right] = 2E(Y, X) \left[ 1_{Y \neq \hat{Y}_{\text{Bayes}}(X)} \right]
\end{align*}
\]

2.2 Nonlinear classification with kernels

2.2.1 Introduction

Linear classifier

\[
\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \theta^T X_i) + \frac{\lambda}{2} \|\theta\|^2
\]

The problem of linear classification is to find a linear decision function, that separates the training data with a hyperplane. In some cases a nonlinear decision function can be better suited to separate the training data.

First idea: transform \(X\) with a nonlinear function \(\varphi: \mathbb{R}^p \rightarrow \mathbb{R}^d\) where \(d \neq p\)

\[
\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \theta^T \varphi(X_i)) + \frac{\lambda}{2} \|\theta\|^2
\]

Q: How to choose \(\varphi\)?

Second idea

\[
\min_{f \in F} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i)) + \lambda \Omega(f)
\]
F is set of nonlinear functions

Q1: What is $\Omega$?

Q2: How do I solve equation 1?

Case of parameterized functions s.t. $F : \{f^\theta, \theta \in \mathbb{R}^p\}$

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f^\theta(X_i)) + \lambda \Omega(f^\theta)$$

Obviously this does not solve Q1, it also does not solve Q2, the problem might be non-convex. One solution to Q1 and Q2: “kernels”

- Extend linear machine learning to non-linear settings, without losing any good properties
- Do not require the $X_i$ to be in $\mathbb{R}^p$, you just need the $X$ to be in some set $\mathcal{X}$

Example $\mathcal{X} =$

$$\begin{cases} 
- \text{graphs} \\
- \text{DNA sequences} \\
- \text{time} \\
- \text{string} \\
- \text{groups}
\end{cases}$$

the only downside is the $O(M^2)$ complexity with the amount of data ($M$ is number of training points).

Useful resources:

- “Machine learning with Kernel methods” course of Jean-Phillipe VERT (http://cbio.ensmp.fr/~jvert/teaching/)

2.2.2 RKHS (Reproducing Kernel Hilbert Space) and kernels

Idea:

- Instead of working with $\mathcal{X}$, work “implicitly” with $\varphi(X)$ in a Hilbert space $\mathcal{H}$
- Reformulate learning problem by “involving” pairwise comparisons between the $X_i$'s
Example For $n = 3$ (3 training points in $\mathcal{X}$), we can define a similarity measure $K$:

$$K = \begin{pmatrix}
1 & 0.6 & 0.1 \\
0.6 & 1 & 0.2 \\
0.1 & 0.2 & 1
\end{pmatrix} \in \mathbb{R}^{m \times m}$$

Define a “comparison function” $K$, called a kernel

- + methods “blind” to the type of data
- + $K$ can be non-linear
- + $K$ will be “plugged” in many algorithms
- - $K$ has $m^2$ entries

Definition (Semi) positive definite kernel or (p.d. kernel) on some set $\mathcal{X}$: $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is p.d. iff

- it is symmetric $K(X, X') = K(X', X)$ $\forall (X, X') \in \mathcal{X} \times \mathcal{X}$
- $\forall (X_1, \ldots, X_N) \in \mathcal{X}^N$ and $(a_1, \ldots, a_n) \in R^N$ then
  \[\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j K(X_i, X_j) \geq 0\] 
  or $a^T K_n a \geq 0$ where $a = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$ and $K_N = [K(x_i, x_j)]_{i,j \in N \times N} \\
  or $K_n$ is semi-positive definite

Motivation Theorem (Aronszajn, 1950)\[1]\]

$K$ is a p.d. kernel iff there exists a Hilbert space $\mathcal{H}$ and a mapping $\varphi : \mathcal{X} \to \mathcal{H}$ such that

$$\forall (x, x') \in \mathcal{X} \times \mathcal{X} \quad K(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$$

Definition of spaces

- Euclidean spaces: Vector space of finite dimension + inner product
  - inner product: bilinear + symmetric + $\langle x, x \rangle \geq 0$ iff $x \neq 0$
- pre-Hilbert: properties of Euclidean space + possibly infinite dimension
• Hilbert: pre-Hilbert + complete
  complete: all Cauchy sequences converge in the space

Cauchy sequence: \((u_m)_{m \geq 0}\) is Cauchy if
\[
\lim_{m \to +\infty} \sup_{p,q \geq m} |u_p - u_q| = 0
\]

Example

• linear Kernel \(X = \mathbb{R}^d\)
  \[-\quad K(x, x') = x^T x' = x'^T x = K(x'x)
  \]
  \[-\quad \text{Consider } x_1, \ldots, x_n \in \mathbb{R}^p \text{ and } (a_1, \ldots, a_n) \in \mathbb{R}^n \text{ and } X = [x_1, \ldots, x_n] \in \mathbb{R}^{p \times n}
  \]
  \[
  \sum_{j=1}^{n} \sum_{i=1}^{n} a_i a_j K(x_i, x_j) \\
  = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j x_i^T x_j \\
  = \left( \sum_{i} a_i x_i \right)^T \left( \sum_{j} a_j x_j \right) \\
  = \langle Xa, Xa \rangle \\
  = \|Xa\|_2^2 \geq 0
  \]

• Polynomial kernel \(K(x, x') = (x^T x')^d\) for \(x \in \mathbb{R}^p\)
  \[-\quad \text{proof for } d=2:\]
  \[-\quad * \text{ symmetric is obvious}
  \]
  \[-\quad * \text{Consider } x_i, \ldots, x_n \in \mathbb{R}^p \text{ and } a = (a_1, \ldots, a_n) \in \mathbb{R}^n
  \]
  \[
  \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \left( x_i^T x_j \right)^2 \\
  \text{trace}(x_i^T x_j x_j^T x_i) \\
  = \text{trace}(x_j^T x_i x_i^T x_j) \\
  \text{trace} \left( \sum_{j=1}^{n} a_j x_j x_j^T \right) \left( \sum_{i=1}^{n} a_i x_i x_i^T \right) \\
  = \sum_{i=1}^{n} a_i x_i x_i^T \sum_{j=1}^{n} a_j x_j x_j^T \geq 0
  \]
it turns out, that \((A, B) \rightarrow \text{trace}(A^T B) = \sum_{i,j} A_{ij} B_{ij}\) is an inner product, the norm associated with it is called the Frobenius norm \(\| \cdot \|_F\).

- Proof of Aronszjan for finite set \(X = \{x_1, \ldots, x_n\}\) \(K\) is p.d. kernel

\[
K_n = [K(x_i, x_j)]_{(i,j) \in \mathbb{R}^n \times n}
\]

\(K_n\) is \(K_n = US^2U^T\) s.t. \(U^T U = \sum_{k=1}^{n} s_k u_k u_k^T\) where \(U = [u_1, \ldots, u_n]\)

\(U\) contains the eigen vectors of \(K_n\), the corresponding eigen values are non-negative because of the p.d. property.

\[
K(x_i, x_j) = \sum_{k=1}^{n} s_k^2 u_k[i] u_k[j] = \langle \varphi(x_i), \varphi(x_j) \rangle \quad \text{where} \quad \varphi(x_i) = \begin{pmatrix} s_1 u_1(i) \\ \vdots \\ s_k u_k(i) \\ \vdots \\ s_n u_n(i) \end{pmatrix} \in \mathbb{R}^n
\]

**Definition RKHS** Let \(\mathcal{X}\) be a set and \(\mathcal{H} \subset \mathbb{R}^\mathcal{X}\) be a class of functions forming a Hilbert space with inner-product \(\langle \cdot, \cdot \rangle_{\mathcal{H}}\)

\(K : \mathcal{X}^2 \rightarrow \mathbb{R}\) is called a reproducing kernel for \(\mathcal{H}\) iff

- \(\mathcal{H}\) contains the functions \(K_x : \mathcal{X} \mapsto \mathbb{R}, K_x : t \mapsto K(x, t) \forall x \in \mathcal{X}\)

- for all \(x \in \mathcal{X}\) and \(f \in \mathcal{H}\) (f: decision function, non-linear, but linear in Hilbert space), then

\[f(x) = \langle f, K_x \rangle_{\mathcal{H}}\]

(Reproducing property)

**Intuition** Spoiler: What’s going to happen next

We are going to consider

\[
\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(x_i)) + \frac{\lambda}{2} \|f\|^2_{\mathcal{H}} \quad (2)
\]
We will show that if $K$ is reproducing for $\mathcal{H}$, then there exists a solution of equation 2 that is a linear combination of $K_{x_i}$,

$$\exists \alpha \in \mathbb{R}^n \text{ s.t. } f = \sum_{i=1}^{n} \alpha K_{x_i}$$

We notice that

$$\|f\|^2_{\mathcal{H}} = \langle f, f \rangle_{\mathcal{H}}$$

$$= \left\langle \sum_{i=1}^{n} \alpha_i k_i, \sum_{j=1}^{n} a_j k_j \right\rangle_{\mathcal{H}}$$

$$= \sum_{i,j} \alpha_i \alpha_j \langle K_{x_i}, K_{x_j} \rangle_{\mathcal{H}}$$

$$= \sum_{i,j} \alpha_i \alpha_j K_{x_i}(y_j)$$

$$= \alpha^T K_n \alpha$$

$$f(x_i) = \langle f, K_{x_i} \rangle_{\mathcal{H}} = \left\langle \sum_{j=1}^{n} \alpha_j K_{x_j}, K_{x_i} \right\rangle_{\mathcal{H}}$$

$$= \sum_{j=1}^{n} \alpha_j K(x_j, x_i)$$

$$= [K_n \alpha]_i$$

**Theorem** A function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is p.d. iff it is a reproducing kernel of a Hilbert space $\mathcal{H}$.

**Theorem** If $\mathcal{H}$ is a RKHS, it has a unique kernel. Conversely, a function $K$ is reproducing for at most one Hilbert space

**Example of RKHS** linear kernel

Q: What is $\mathcal{H}$?

Candidate: $\mathcal{H}_0 = \{f_x : t \mapsto x^T t; x \in \mathbb{R}^p\}$

definition of inner-product: $\langle f_x, f_y \rangle_{\mathcal{H}_0} = x^T y$
∀x' ∈ Rp and f_x ∈ H_0

f_x(x') = x^T x' = ⟨f_x, f_x⟩_{H_0} = ⟨f_x, K_{x'}⟩_{H_0}

→ therefore: H = H_0

for K(x, x') = (x^T x')^2, H = \{t ↦ t^T Z t, Z symmetric matrix \}

References

