Institute of Neural Information Processing | Ulm University

INSTANCE-BASED LEARNING

Dr. Sebastian Gottwald

MOTIVATION

BASIC IDEA

- **Given:** Data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ of patterns $x_i \in \mathcal{X}$ and targets $y_i \in \mathcal{Y}$.
- **Goal:** Predict the target y of a new pattern $x \in \mathcal{X}$.
- Need **additional structure** to compare x to the known information:
  - Parametrized models use a **loss function** defined on $\mathcal{Y}$ to **compare outputs** during training (and use the error to adapt the model parameters, e.g. gradient-descent).
  - (most) instance-based models use a **kernel function** defined on $\mathcal{X}$ to **compare inputs**

Note: Here, a kernel is a non-negative function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+, (x, x') \mapsto k(x, x')$ that is used to measure similarity of x and x', but often kernels are required to satisfy additional constraints such as positive definiteness or symmetry (we will see this later).

EXAMPLE 1: KERNEL REGRESSION (NADARAYA-WATSON MODEL)

See notes for details.

- Consider dataset $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$ as a sample from a joint distribution $\mathbb{P}(X, Y)$.
- Approximate the joint density by $p(x, y) := \frac{1}{N} \sum_{n=1}^N f(x - x_n) g(y - y_n)$.
- Define the regression function to be $y(x) := \mathbb{E}_{p(Y|X=x)}[Y]$.
- A short calculation shows that

$$y(x) = \frac{\int y\, p(x, y) dy}{\int p(x, y) dy} = \cdots = \sum_n y_n \underbrace{\frac{f(x - x_n)}{\sum_m f(x - x_m)}}_{=: k_D(x, x_n)} = \sum_n y_n k_D(x, x_n)$$

Note: $k_D(x, x_n) \in [0, 1]$ and $\sum_n k_D(x, x_n) = 1$, i.e. $q(n) := k_D(x, x_n)$ can be viewed as a probability distribution over the patterns x_n, favoring those x_n that are considered "similar" to x, and thus $y(x)$ is the average of y_n wrt. q.

INNER PRODUCT KERNELS

If a kernel k can be written as an **inner product** on some space $\mathcal{H}$, a so-called *feature space*, in the sense that

$$k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$$

for some mapping $\phi : \mathcal{X} \to \mathcal{H}$, a so-called **feature map**, then k is called an **inner product kernel** (often the prefix *inner product* is dropped!).

The image $\phi(x) \in \mathcal{H}$ of a pattern x under ϕ is then called a *feature vector*, and its components are called *features*.

WHY USE INNER PRODUCTS?

Inner products can serve as **similarity measures** in vector spaces: For $x, x' \in \mathbb{R}^d$, we have

$$\langle x, x' \rangle = \|x\| \, \|x'\| \, \cos \phi(x, x')$$

where $\phi(x, x')$ is the angle between x and x'.

- If $\|x\| = \|x'\| = 1$, then $\langle x, x' \rangle \in [-1, 1]$ is maximal if $\cos \phi(x, x') = 1$, i.e. if x and x' point in the same direction.
- Hence, $\langle x, x' \rangle$ is a good similarity measure, if the length (or *magnitude*) of the vectors are not informative.
- E.g. when the vectors are normalized in some way, so that only the proportions between the features are relevent, not their total value.
- Or when having decision hyperplanes through the origin, so that only the angle is used as a criterion.

Note: The angle between two vectors can also be defined in arbitrary/infinite dimensional inner product spaces by the above equality (due to the *Cauchy-Schwarz inequality*, c.f. next section).

WHY USE FEATURE MAPS?

- If $\mathcal{X}$ is a set without vector space structure (e.g. words), then a feature map ϕ **embeds** $\mathcal{X}$ into an inner product space, where the inner product allows to measure similarity.
- Even if $\mathcal{X}$ is already a vector space with an inner product, it might not measure the right notion of similarity for a given problem.

WHY USE INNER PRODUCT KERNELS?

The features might live in a very high (maybe even infinite) dimensional space, but the kernel could have a closed form that does not require the explicit calculation of the features.

RULE OF THUMB: FEATURE MAPS VS KERNELS

- **Kernels** have an advantage when the feature space is high dimensional
- **Feature maps** are better if the number of samples is very large

SOME FEATURE MAPS AND THEIR KERNELS

Feature map	$\Rightarrow$	**Kernel**
$\phi : \mathbb{R}^d \to \mathbb{R}^d, x \mapsto x$		$k(x, x') = \langle x, x' \rangle_{\mathbb{R}^d} = \mathbf{x}^T \mathbf{x}'$
$\phi : \mathbb{R}^d \to \mathbb{R}^{d^2}, x \mapsto (x_i x_j)_{i,j=1}^d$		$k(x, x') = \left(\langle x, x' \rangle_{\mathbb{R}^d} \right)^2$
$\phi : \{0, 1, 2\} \to [0, 1], x \mapsto p(x)$		$k(x, x') = p(x)p(x')$
$\phi : 2^\Omega \to L^\infty(\Omega), A \mapsto \mathbb{1}_A - P(A)$		$k(A, B) = P(A \cap B) - P(A)P(B)$

The converse: How do we know that a kernel, e.g. $f(x, x') = e^{-\|x-x'\|^2}$, is an inner product kernel (i.e. can be written as an inner product of $\phi(x)$ and $\phi(x')$ for some ϕ)?

Answer: Hilbert space theory (next section).

REPRODUCING KERNEL HILBERT SPACES

VECTOR SPACES

A *vector space* (or *linear space*) V consists of elements v (called *vectors*) that can be added ($v + w \in V$, if $v, w \in V$) and multiplied by scalars ($\alpha v \in V$ if $\alpha \in \mathbb{R}, v \in V$). Examples include

- **Euclidean spaces** $\mathbb{R}^d$: $\alpha x + y \in \mathbb{R}^d$, where $(\alpha x + y)_i := \alpha x_i + y_i$ (elementwise)
- **Sequence spaces:** $\alpha(x_n)_n + (y_n)_n := (\alpha x_n + y_n)_n$ (elementwise), e.g. bounded sequences ℓ^∞, summable sequences ℓ^1, square-summable sequences $\ell^2, \ldots$.
- **Function spaces:** $(\alpha f + g)(x) := \alpha f(x) + g(x)$ (pointwise), e.g. continuous functions $C([a, b])$ on an interval $[a, b]$, continuously differentiable functions $C^1((a, b))$, square integrable functions $L^2(\mathbb{R})$ on $\mathbb{R}, \ldots$

Note: For the purpose of this lecture, we assume that $L^2(\mathbb{R})$ consists of functions. Rigorously, one has to consider equivalence classes of functions that are equal *almost everywhere*, which means that $f, g \in L^2(\mathbb{R})$ are considered the same even if $f(x) \neq g(x)$ on a set of measure 0 ($A \subset \mathbb{R}$ has measure 0 if $\int_A dx = 0$, e.g. $A = \{x\}\ \forall x \in \mathbb{R}$).

INNER PRODUCT SPACES

An *inner product space* is a vector space V together with an *inner product* $\langle \cdot, \cdot \rangle$, which (in the real case) is a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ that is **symmetric**, **linear** in both entries, and **positive definite** ($\langle x, x \rangle > 0$ if $x \neq 0$).

Examples:

- Euclidean spaces $(\mathbb{R}^d, \langle \cdot, \cdot \rangle_{\mathbb{R}^d})$, where $\langle x, y \rangle_{\mathbb{R}^d} = \sum_{i=1}^d x_i y_i$ for $x, y \in \mathbb{R}^d$.
- Sequence spaces, e.g. $(\ell^2, \langle \cdot, \cdot \rangle_{\ell^2})$, where $\langle x, y \rangle_{\ell^2} = \sum_{i=1}^\infty x_i y_i$ for $x, y \in \ell^2$.
- Function spaces, e.g. $(L^2(\mathbb{R}), \langle \cdot, \cdot \rangle_{L^2})$, where $\langle f, g \rangle_{L^2} = \int_{\mathbb{R}} f(x) g(x)\, dx$.

INDUCED NORM

An inner product space V is an example of a **normed space** with norm $\|v\| = \sqrt{\langle v, v\rangle}$ for all $v \in V$. A norm measures the *length* of a vector v, and therefore introduces a notion of *distance* into V by $d(v, w) = \|v - w\|$ (a so-called *metric*), which, in turn, implies a notion of convergence (a *topology*).

Examples of **norms** that are **induced by inner products**:

- Euclidean norm: $\|x\| = \sqrt{\sum_{i=1}^{d} x_i^2} = \sqrt{\langle x, x\rangle_{\mathbb{R}^d}}$ for $x \in \mathbb{R}^d$ (analogous for ℓ^2)
- Function norm in L^2: $\|f\| = \sqrt{\int |f(x)|^2 dx} = \langle f, f\rangle_{L^2}$ for $f \in L^2(\mathbb{R})$.

Examples of **norms** that do **not come from inner products**:

- ℓ^p and L^p norms for $p \neq 2$: $\|x\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$, $\|f\|_p = \left(\int |f(x)|^p dx\right)^{1/p}$
- the supremum norms $\|x\|_\infty = \sup_i |x_i|$ and $\|f\|_\infty = \sup_x |f(x)|$.

CAUCHY-SCHWARZ INEQUALITY

Theorem: *For all elements $v, w \in V$ of an inner product space V, we have*

$$|\langle v, w \rangle| \leq \|v\| \|w\|.$$

- In $\mathbb{R}^d$, this can be seen as a consequence of $\langle x, y \rangle = \|x\| \|y\| \cos \theta$. In fact, it justifies the **definition of an angle** between elements of arbitrary inner product spaces.
- It implies the **triangle inequality**, $\|v + w\| \leq \|v\| + \|w\|$, in any inner product space (see exercises).
- It is very useful to show implications like $x, y \in \ell^2 \Rightarrow xy \in \ell^1$ (see exercises), which is why it appears all over Analysis.

HILBERT SPACES

A *Hilbert space* $\mathcal{H}$ is an inner product space with the additional property that all sequences $(x_n)_n$ in $\mathcal{H}$ whose elements are eventually arbitrarily close to each other (so-called **Cauchy sequences**) **do converge** to elements in $\mathcal{H}$. Normed spaces with this property are known as being *complete*. Examples:

- **Hilbert Spaces:** The inner product spaces from the previous slides ($\mathbb{R}^d, \ell^2, L^2$)
- **Non-complete** inner product spaces: Rational numbers $\mathbb{Q}$ (equipped with product of numbers), $C([a, b])$ equipped with $\langle \cdot, \cdot \rangle_{L^2([a,b])}$.

Note: Any (non-complete) inner product space can be uniquely completed to a Hilbert space by simply including all limits of Cauchy sequences as elements of the space, e.g. the completion of $\mathbb{Q}$ is $\mathbb{R}$, the completion of $(C([a, b]), \langle \cdot, \cdot \rangle_{L^2})$ is $L^2([a, b])$.

DUAL SPACES

The (topological) *dual* X^* of a (topological) space X consists of all **continuous linear maps** (so-called *functionals*) $\phi : X \to \mathbb{R}$. Examples:

1. Inner product by a fixed vector $a \in \mathbb{R}^n$, i.e. $\phi : \mathbb{R}^n \to \mathbb{R}$ with $\phi(x) = \langle a, x \rangle_{\mathbb{R}^n}$.
2. Summation on ℓ^1 against a fixed bounded sequence $(y_n)_n$ ($\exists C$ s.th. $|y_n| \leq C \, \forall n$), i.e. $\phi : \ell^1 \to \mathbb{R}$ with $\phi(x) = \sum_n y_n x_n$.
3. Integration on $L^2(\mathbb{R})$ against a fixed function $g \in L^2(\mathbb{R})$, i.e. $\phi : L^2(\mathbb{R}) \to \mathbb{R}$ with $\phi(f) := \int_{\mathbb{R}} g(x) f(x)\, dx$.

Note: 1. and 3. are examples of the general fact that, in any Hilbert space $\mathcal{H}$, the inner product against a fixed element $y \in \mathcal{H}$, i.e. $\phi(x) = \langle y, x \rangle_{\mathcal{H}}$, defines a continuous linear functional $\phi \in \mathcal{H}^*$ (exercise).

RIESZ REPRESENTATION THEOREM

The following theorem shows that the dual $\mathcal{H}^*$ of a Hilbert space $\mathcal{H}$ can be identified with the Hilbert space itself.

Theorem (Riesz): *For every continuous linear functional $\phi : \mathcal{H} \to \mathbb{R}$, there exists a unique element $g_\phi \in \mathcal{H}$ such that*

$$\phi(f) = \langle g_\phi, f \rangle \quad \forall f \in \mathcal{H}$$

Since, the converse is also true (see comment on the previous slide), the mappings $\phi \mapsto g_\phi$ and $g \mapsto \langle g, \cdot \rangle$ are inverses of each other and allow to **identify $\mathcal{H}^*$ with $\mathcal{H}$**.

Note: For the rigorous identification of $\mathcal{H}$ and $\mathcal{H}^*$ one also has to think of how the distance measure given by the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ transforms under the bijection (we are not doing this here).

EXAMPLE: EVALUATION FUNCTIONALS

For $x \in \mathcal{X}$, an *evaluation functional* $\delta_x : \mathcal{H} \to \mathbb{R}$ on a Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$ is defined by

$$\delta_x(f) := f(x).$$

- δ_x is always linear by definition: $\delta_x(\alpha f + g) = \alpha f(x) + g(x) = \alpha\delta_x(f) + \delta_x(g)$
- δ_x is not necessarily continuous, e.g. in $L^2(\mathbb{R})$, even if $\|f - f_n\| \to 0$, the value $f_n(x)$ can be arbitrarily far away from $f(x)$ for any $n \in \mathbb{N}$ ($\{x\}$ has measure 0).
- In $\mathbb{R}^d$, evaluation functionals δ_i map vectors x to single entries x_i. Thus,

$$\delta_i(x) = x_i = \sum_{j=1}^{d} \delta_{ij} x_j = \langle (\delta_{ij})_{j=1}^{d}, x \rangle$$

  in particular, δ_i is continuous, and the element $y \in \mathbb{R}^d$ that is guaranteed to exist by the Riesz representation theorem in this case is $y = (\delta_{ij})_{j=1}^{d}$.

REPRODUCING KERNEL HILBERT SPACES

Let $\mathcal{H}$ be Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ such that the evaluation functionals $\delta_x : \mathcal{H} \to \mathbb{R}, f \mapsto f(x)$ are continuous, i.e. $\delta_x \in \mathcal{H}^*$, for all $x \in \mathcal{X}$. By Riesz' representation theorem, for every $x \in \mathcal{X}$ there exists an element (i.e. a function)

$$k_x \in \mathcal{H} \quad s.th. \quad f(x) = \langle k_x, f \rangle \quad \forall f \in \mathcal{H}. \tag{$*$}$$

Any function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $k_{x'}(x) := K(x, x')$ satisfies $(*)$ is called a *reproducing kernel for* $\mathcal{H}$, and $\mathcal{H}$ is called a *reproducing kernel Hilbert space* (RKHS) if it has a reproducing kernel (e.g. $\mathbb{R}^d$, ℓ^2, not L^2).

Note: The argument leading to $(*)$ shows that any Hilbert space of functions with continuous evaluation functionals ($\delta_x \in \mathcal{H}^*$) is an RKHS. The converse is also true: if $\mathcal{H}$ has a reproducing kernel K, then δ_x is continuous, since $\delta_x(f) = \langle K(\cdot, y), f \rangle$ and $\langle \cdot, \cdot \rangle$ is continuous in both entries.

PROPERTIES OF REPRODUCING KERNELS

Let K be a reproducing kernel for $\mathcal{H}$, then

1. K is **unique** (as a reproducing kernel of $\mathcal{H}$).

   Proof: Choose $f = K_1(\cdot, x') - K_2(\cdot, x')$ in $\langle K_1(\cdot, x), f\rangle - \langle K_2(\cdot, x), f\rangle = f(x) - f(x) = 0$.

2. K is an **inner product kernel**: $K(x, x') = \langle K(\cdot, x), K(\cdot, x')\rangle$

   Proof: Choose $f = k_{x'} = K(\cdot, x')$ in $(*)$.

3. K is **symmetric**: $K(x, x') = K(x', x)$

   Proof: This directly follows from 2. and the symmetry of inner products.

4. K is **positive semi-definite**, i.e. $K_{ij} := K(x_i, x_j)$ defines a positive semi-definite matrix for any finite set $\{x_1, \dots, x_n\} \subset \mathcal{X}$, i.e. $\sum_{i,j=1}^{n} c_i c_j K_{ij} \geq 0 \; \forall c \in \mathbb{R}^n$.

   Proof: $\sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) = \langle \sum_{i=1}^{n} c_i K(\cdot, x_i), \sum_{j=1}^{n} c_j K(\cdot, x_j)\rangle = \| \sum_{i=1}^{n} c_i K(\cdot, x_i)\|^2 \geq 0$

NATIVE SPACES

Theorem (Moore-Aronszajn): *A symmetric function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel for a unique Hilbert space $\mathcal{H}$ of functions on $\mathcal{X}$ if K is positive semi-definite.*

Sketch of proof:

- Consider the inner product space $V := \mathrm{span}\{K(\cdot, x) : x \in \mathcal{X}\}$ of all finite linear combinations $\sum_{i=1}^{n} \alpha_i K(\cdot, x_i)$ with inner product

$$\left\langle \sum_i \alpha_i K(\cdot, x_i), \sum_j \beta_j K(\cdot, x_j) \right\rangle_V := \sum_{i,j} \alpha_i \beta_j K(x_i, x_j)$$

- Check the reproducing property $(*)$: $f(x) = \langle K(\cdot, x), f \rangle$ for all $f \in V$.

- Define $\mathcal{H}$ as the completion of V (the reproducing property still holds).

Note: Some books (Wendland, Fasshauer) require K to be positive definite in order to get a positive definite inner product, even though semi-definite is enough because one can show $|f(x)|^2 = |\langle K(\cdot, x), f \rangle|^2 \leq K(x, x) \langle f, f \rangle_V$, so $\langle f, f \rangle_V = 0$ implies $f = 0$ (see e.g. Mohri et al. Sect. 5.2.2, or Schölkopf et al., Sect. 1.2).

REPRESENTER THEOREM

Consider a supervised learning problem for given data $\{(x_i, y_i)\}_{i=1}^{N} \subset \mathcal{X} \times \mathbb{R}$. Let l_f be a loss function with respect to a model $f : \mathcal{X} \to \mathbb{R}$, e.g. $l_f(x, y) = (y - f(x))^2$. Consider the regularized optimization problem

$$\min_{f:\mathcal{X}\to\mathbb{R}} \frac{1}{N} \sum_{i=1}^{N} l_f(x_i, y_i) + \lambda\, g(\|f\|) \qquad (**)$$

where $g : \mathbb{R}_+ \to \mathbb{R}$ is a strictly monotonically increasing function, e.g. $g(t) = t^2$, and $\|f\|$ is some a function norm.

Theorem: *If the minimization in $(**)$ is restricted to an RKHS $\mathcal{H}$ with kernel K and $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle_{\mathcal{H}}}$, then each minimizer of $(**)$ admits a representation of the form*

$$f(x) = \sum_{i=1}^{N} \alpha_i\, K(x_i, x)$$

where $\alpha = (\alpha_1, \ldots, \alpha_N) \in \mathbb{R}^N$ is the only degree of freedom that is left.

KERNEL MACHINES

LINEAR SUPPORT VECTOR MACHINE

Consider a binary classification problem for a dataset $\{(x_i, y_i)\}_{i=1}^N, y_i \in \{-1, 1\}$.

- *Parametrized hyperplane* $h_{w,b} := \{\xi | \langle w, \xi \rangle + b = 0\}$
- **Decision function** $f_{w,b}(x) := \operatorname{sgn}(\langle w, x \rangle + b) \in \{-1, 1\}$
- *Margin* $m_{w,b}$:= **distance of** $h_{w,b}$ **to closest points** $= \pm\left(\left\langle \frac{w}{\|w\|}, x_\pm^* \right\rangle + \frac{b}{\|w\|}\right)$
- *Scaling invariance:* $h_{w,b} = h_{\alpha w, \alpha b}$ and $m_{w,b} = m_{\alpha w, \alpha b}$ for any $\alpha \neq 0$.
- *Scaling trick (canonical form):* Rescale w such that $\|w\| = \frac{1}{m_{w,b}}$ (dep. on w and b), resulting in $\langle w, x_\pm^* \rangle + b = \pm 1$ and $m_{w,b} = \frac{1}{\|w\|}$.
- *Max. margin classifier* (**linear SVM**): $\min_{w,b} \frac{1}{2}\|w\|^2$ s.t. $y_i(\langle w, x_i \rangle + b) \geq 1 \ \forall i$.

See notes for details.

TRANSFORMING CONSTRAINED TO UNCONSTRAINED OPTIMIZATION

A **constrained** optimization problem

$$\min_\omega f(\omega) \quad \text{subject to } c_i(\omega) \leq 0 \; \forall i \in 1, \ldots, N \tag{$*$}$$

can be formally translated to the **unconstrained** problem $\inf_\omega F(\omega)$ where

$$F(\omega) = \begin{cases} f(\omega) & \text{if } c_i(\omega) \leq 0 \; \forall i \in \{1, \ldots, N\} \\ \infty & \text{otherwise} \end{cases}$$

Main example: $F(\omega) = \sup_{\lambda_i \geq 0} \mathcal{L}(\omega, \lambda)$ with the **Lagrangian**

$$\mathcal{L}(\omega, \lambda) := f(\omega) + \sum_{i=1}^{N} \lambda_i c_i(\omega),$$

so that $(*)$ can be written as $\inf_\omega \sup_{\lambda_i \geq 0} \mathcal{L}(\omega, \lambda)$.

DUALITY IN CONSTRAINED OPTIMIZATION

So far (trivial): $\min_\omega f(\omega)$ s.t. $c_i(\omega) \leq 0\ \forall i \in 1, \ldots, N \iff \inf_\omega \sup_{\lambda_i \geq 0} \mathcal{L}(\omega, \lambda)$

Strong duality: Can we **interchange the** $\sup$ **and** $\inf$ **operators**?
More precisely, strong duality means that, if $g(\lambda) := \inf_\omega \mathcal{L}(\omega, \lambda)$, then

$$\underbrace{\inf_\omega F(\omega)}_{\textit{Primal Problem}} = \underbrace{\sup_{\lambda_i \geq 0} g(\lambda)}_{\textit{Dual Problem}}$$

Examples of sufficient conditions for strong duality (there are many!):

- f and all c_i are affine functions (linear optimization problem)
- f is **convex and all** c_i **are affine** (variant of *Slater's condition*)
- f and all c_i are convex and continuous on a compact and convex domain (*minimax thm.*)

Theorem (Bazaraa et al. 2006, Thm. 6.2.5): *ω^* and λ^* are solutions of the primal and dual problems, respectively, and strong duality holds, if and only if (ω^*, λ^*) is a saddle point of $\mathcal{L}$, i.e. $\mathcal{L}(\omega^*, \lambda) \leq \mathcal{L}(\omega^*, \lambda^*) \leq \mathcal{L}(\omega, \lambda^*)$ for all ω, λ.*

KARUSH-KUHN-TUCKER (KKT) CONDITIONS

Assume that strong duality holds for a pair ω^*, λ^* (and that f, c_i are differentiable), then

- $c_i(\omega^*) \leq 0$ and $\lambda_i^* \geq 0$ for all $i = 1, \ldots, N$ (*feasability*)
- $\frac{\partial \mathcal{L}}{\partial \omega}(\omega^*, \lambda^*) = 0$ (*stationarity* of $\mathcal{L}(\omega, \lambda^*)$ at $\omega = \omega^*$)
- $\lambda_i^* c_i(\omega^*) = 0$ for all $i = 1, \ldots, N$ (*complementary slackness*)

These are known as the *Karush-Kuhn-Tucker (KKT) conditions*.

Theorem (see e.g. Chi et al. 2017, Sect. 9.5):
(i) *If strong duality holds, then the above conditions follow for a pair ω^*, λ^* of solutions.*
(ii) *For convex problems with strong duality (e.g. Slater's condition holds), the KKT conditions are also sufficient for ω^*, λ^* being solutions for the primal and dual problems, respectively.*

Note: One can find many regularity conditions in the optimization literature (so-called *constraint qualifications*, e.g. Peterson, 1973) under which the KKT conditions are necessary, but one does not necessarily have strong duality.

DUAL PROBLEM FOR LINEAR SVM

- **Primal problem:** $\min_{w,b} \frac{1}{2}\|w\|^2$ subject to $1 - y_i(\langle w, x_i \rangle + b) \leq 0 \; \forall i \in \{1, \dots, N\}$
- **Lagrangian:** $\mathcal{L}(w, b, \lambda) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^N \lambda_i \big(1 - y_i(\langle w, x_i \rangle + b)\big)$
- Since $f(w) = \frac{1}{2}\|w\|^2$ is convex, and the constraints are affine (*variant of Slater's condition*), we have **strong duality**. In particular, the KKT conditions are necessary and sufficient. Moreover, we can maximize $g(\lambda) := \min_{w,b} \mathcal{L}(w, b, \lambda) = \mathcal{L}(w^*(\lambda), b^*(\lambda), \lambda)$ over $\lambda_i \geq 0$, where $w^*(\lambda)$ and $b^*(\lambda)$ satisfy

$$\underbrace{\frac{\partial \mathcal{L}}{\partial w_i}(w^*(\lambda), b^*(\lambda), \lambda) = 0}_{w^*(\lambda)=\sum_{i=1}^N \lambda_i y_i x_i}\,, \qquad \underbrace{\frac{\partial \mathcal{L}}{\partial b}(w^*(\lambda), b^*(\lambda), \lambda) = 0}_{\sum_{i=1}^N \lambda_i y_i=0}$$

$\Rightarrow$ **Dual problem** (see notes for details): $\max_{\lambda_i \geq 0} g(\lambda)$ subject to $\sum_{i=1}^N \lambda_i y_i = 0$, where

$$g(\lambda) = \mathcal{L}(w^*(\lambda), b^*(\lambda), \lambda) = \sum_{i=1}^N \lambda_i - \frac{1}{2} \sum_{i,j=1}^N \lambda_i \lambda_j y_i y_j \langle x_i, x_j \rangle$$

SUPPORT VECTORS

By complementary slackness, $\lambda_i^*(1 - y_i(\langle w^*, x_i \rangle + b^*)) = 0$ for $i = 1, \ldots, N$, i.e.

$$\lambda_i^* = 0 \quad \text{or} \quad y_i(\langle w^*, x_i \rangle + b^*) = 1 \tag{$*$}$$

This means that in the linear combination $w^*(\lambda^*) = \sum_i \lambda_i^* y_i x_i$ only those patterns x_i contribute that satisfy the constraint as an equality (they are on the margin!) known as *support vectors*. In particular, **all other patterns have no influence** on the optimal hyperplane.

DECISION FUNCTION

Plugging in the expression for $w^*(\lambda^*)$ into the decision function $f_{w,b}$ of the linear SVM, we obtain

$$f_{w^*(\lambda^*),b^*}(x) = \operatorname{sgn}\Big(\sum_{i=1}^N \lambda_i^* y_i \langle x_i, x \rangle + b^* \Big)$$

where λ^* is given by the dual problem, and, due to $(*)$, $b^* = y_j - \sum_{i=1}^N \lambda_i^* y_i \langle x_i, x_j \rangle$ for all j with $\lambda_j^* > 0$ (e.g. by averaging).

NONLINEAR SVM

In the linear SVM with decision function $x \mapsto \operatorname{sgn}(\sum_{i=1}^{N} \lambda_i^* y_i \langle x_i, x \rangle + b^*)$, a new pattern x is **compared with all support vectors** x_i using $\langle x_i, x \rangle$ as similarity measure and then categorized based on the weighted sum of these similarities.

Above, the dimension of the x_i was arbitrary. Thus we can **replace them by their image under a feature map** $\phi : \mathcal{X} \to \mathcal{H}$ into a feature space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, so that

$$f(x) = \operatorname{sgn}\Big(\sum_{i=1}^{N} \lambda_i^* y_i K(x_i, x) + b^* \Big)$$

where $K(x, x') := \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$ defines an inner product kernel, and λ^* is a solution to the dual problem $\max_{\lambda_i \geq 0} g(\lambda)$ s.t. $\sum_{i=1}^{N} \lambda_i y_i = 0$, where in, analogy to the linear SVM,

$$g(\lambda) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{N} \lambda_i \lambda_j y_i y_j K(x_i, x_j)$$

Note: In the exercises for this section you will use these results to create simulations for linear and nonlinear Support Vector Machines. You can view my implementations here.

EXTENSIONS OF STANDARD SVMS

- *Soft Margin* (in case of overlapping classes due to noisy data): Introduce *slack variables* $\xi_i \geq 0$, relax the constraints to $y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i$, and minimize $C \sum_i \xi_i$ additionally to $\|w\|^2$, where $C > 0$ denotes a trade-off parameter. The corresponding dual problem takes exactly the same form as the hard margin SVM from the previous slides, with the additional constraint that $\lambda_i \leq C$ (so that $\lambda_i \in [0, C]$) $\forall i = 1, \ldots, N$.

- *SVM Regression (linear and kernel regression):* Analogous to soft margins, one introduces *slack variables* $\xi_i, \xi_i^* \geq 0$ and minimizes $\|w\|^2 + C \sum_i (\xi_i + \xi_i^*)$ subject to $f(x_i) - y_i \leq \varepsilon + \xi_i$ and $y_i - f(x_i) \leq \varepsilon + \xi_i^*$ for some $\epsilon > 0$, where $f(x) := \langle w, x \rangle + b$. This can be transformed to a dual problem with Lagrange multipliers λ_i, λ_i^* and a decision function of the form $f(x) = \sum_i (\lambda_i^* - \lambda_i) K(x_i, x) + b$.

THE KERNEL TRICK

Consider a learning algorithm whose prediction function takes the form

$$f(x) = F\Big(\sum_{i=1}^{N} \alpha_i(y_i) K(x_i, x) + b \Big)$$

where K is some inner product kernel. Then we can obtain a new algorithm by simply replacing the kernel K by another inner product kernel K'.

Examples include

- Linear SVM (classification) $\Longrightarrow$ **Nonlinear/kernel SVM**
- Linear SVM Regression $\Longrightarrow$ **Kernel regression**
- Principal component analysis (PCA) $\Longrightarrow$ **Kernel PCA**

INSTANCE-BASED METHODS NOT RELYING ON INNER PRODUCTS

- k *nearest neighbour* classification: Choose the label that is most common under the k nearest neighbours (given some notion of distance).
- *k nearest neighbour* regression: Average the values of the k nearest neighbours.
- *RBF network* regression: $f(x) = \sum_n \alpha_n h(\|x - x_n\|)$ for some localized function h (usually a Gaussian)