Unit 6 Review (Kernel Methods)

- Extending models into **high dimensional feature spaces** through $\phi(\boldsymbol{x}_1)$
  - Example: Ridge Regression $\min_{\boldsymbol{w}} \|\boldsymbol{y}-\boldsymbol{\Phi w}\|^2+\lambda\|\boldsymbol{w}\|^2$ where $\boldsymbol{\Phi}=[\phi(\boldsymbol{x}_1) \quad \cdots \quad \phi(\boldsymbol{x}_N)]^T$

$$\boldsymbol{w}^*=\left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}+\lambda\boldsymbol{I}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{y}=\boldsymbol{\Phi}^{\mathrm{T}}\left(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}+\lambda\boldsymbol{I}\right)^{-1}\mathbf{y}$$

- **Kernel methods** compute the kernel function $K(\boldsymbol{x},\boldsymbol{x}_i)=\phi(\boldsymbol{x})^T\phi(\boldsymbol{x}_i)$ without explicitly evaluating $\phi(\boldsymbol{x})$.

$$\hat{y}(\boldsymbol{x})=\phi(\boldsymbol{x})^T\boldsymbol{w}^*=\phi(\boldsymbol{x})^T\boldsymbol{\Phi}^{\mathrm{T}}\left(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}+\lambda\boldsymbol{I}\right)^{-1}\mathbf{y}=\sum_{i=1}^{N}\alpha_i K(\boldsymbol{x},\boldsymbol{x}_i)$$

$$\alpha_i=\left[(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}+\lambda\boldsymbol{I})^{-1}\mathbf{y}\right]_{\mathrm{i}}=[(\boldsymbol{K}+\lambda\boldsymbol{I})^{-1}\mathbf{y}]_{\mathrm{i}}$$

- **Popular Kernels:**
  - Linear Kerner: $K(\boldsymbol{u},\boldsymbol{v})=\boldsymbol{u}^T\boldsymbol{v}$
  - Monomials of degree q: $K(\boldsymbol{u},\boldsymbol{v})=(\boldsymbol{u}^T\boldsymbol{v})^q$
  - Polynomials of degree q: $K(\boldsymbol{u},\boldsymbol{v})=(1+\boldsymbol{u}^T\boldsymbol{v})^q$
  - Gaussian Radial Kerner: $K(\boldsymbol{u},\boldsymbol{v})=\exp\left(-\frac{\|\boldsymbol{u}-\boldsymbol{v}\|^2}{2\sigma^2}\right)$
- **Kernel SVMs** introduces high-dimensional feature spaces in SVMs using Kernels.

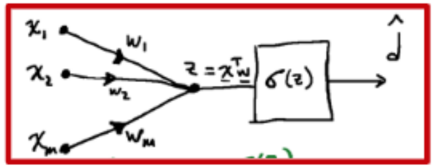
$$\min_{\boldsymbol{w}}\sum_i(1-d_i\phi(\boldsymbol{x}_i)^T\boldsymbol{w})_+ +\lambda\|\boldsymbol{w}\|^2=\min_{\boldsymbol{\alpha}}\sum_{i=1}^{N}\left(1-d_i\sum_{j=1}^{N}\alpha_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)\right)_+ +\lambda\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)$$

- Kernel SVMs optimize for $\boldsymbol{\alpha}$ using gradient descent. They search for a max-margin classifier in the high-dimensional feature space.

Unit 6 Review (Neural Networks)

- The **artificial neuron**: $d = \sigma(\boldsymbol{x}^T \boldsymbol{w})$

  Activation function | Neuron input | Learnable weights



- Common activation functions:

  ReLU: $\sigma(z) = \max(0, z)$ **Logistic**: $\sigma(z) = \frac{1}{1+\exp(-z)}$ **Sign**: $\sigma(z) = sign(z)$

- **Neural networks** $f(\boldsymbol{x})$ are networks of neurons

  Neurons can be both in parallel and in sequence.

- **Training a neural network** means finding the optimal set of weights that minimize a specified loss L.

  Example ridge regression: $\min_{\{\boldsymbol{w}\}} \sum_i (f(x_i; \{\boldsymbol{w}\}) - d_i)^2 - \lambda \|\boldsymbol{P}\|^2$

  Solved by **gradient descent**: $\boldsymbol{w}^{k+1} = \boldsymbol{w}^k - \tau \nabla_w L$

  Back-propagation allows us to efficient compute the gradients of the loss w.r.t. the parameters.