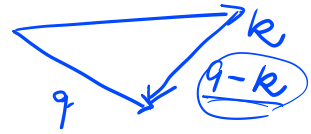
10.2. Attention Pooling by Similarity

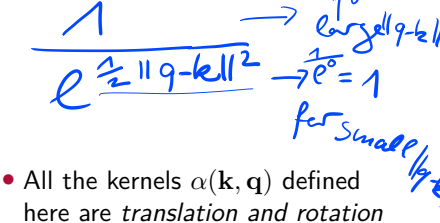
Lecture based on "Dive into Deep Learning" **http://D2L.AI** (Zhang et al., 2020)

Prof. Dr. Christoph Lippert

Digital Health & Machine Learning

Attention by similarity





Nadaraya-Watson estimators rely on a similarity kernel $\alpha(\mathbf{q}, \mathbf{k})$ relating queries $\mathbf{q}$ to keys $\mathbf{k}$. Some common kernels are

$$\alpha(\mathbf{q}, \mathbf{k}) = \exp\left(-\frac{1}{2}\|\mathbf{q} - \mathbf{k}\|^2\right) \quad \text{Gaussian}$$

$$\alpha(\mathbf{q}, \mathbf{k}) = 1 \text{ if } \|\mathbf{q} - \mathbf{k}\| \leq 1 \quad \text{Boxcar}$$

$$\alpha(\mathbf{q}, \mathbf{k}) = \max(0, 1 - \|\mathbf{q} - \mathbf{k}\|) \quad \text{Epanechikov}$$

See for a more extensive review and how the choice of kernels for kernel density estimation.

- All the kernels $\alpha(\mathbf{k}, \mathbf{q})$ defined here are *translation and rotation invariant*.

  i.e., shifting and rotating both $\mathbf{k}$ and $\mathbf{q}$ in the same way

- Different kernels correspond to different notions of range and smoothness.

  e.g., the boxcar kernel only attends to observations within a distance of 1.

Nadaraya Watson Estimators

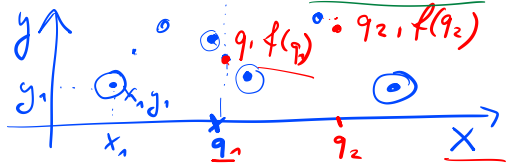
Regression and classification via kernel density estimation

$$\hat{y} = f(\mathbf{q}) = \sum_i \mathbf{v}_i \frac{\alpha(\mathbf{q}, \mathbf{k}_i)}{\sum_j \alpha(\mathbf{q}, \mathbf{k}_j)}.$$

For **regression**:

- observations $(\mathbf{x}_i, y_i)$ (features and labels)
- $\mathbf{v}_i = y_i$ scalars
- $\mathbf{k}_i = \mathbf{x}_i$ are vectors
- query $\mathbf{q}$ denotes the new location where f should be evaluated.

For (multiclass) **classification**, we use one-hot-encoding of y_i to obtain $\mathbf{v}_i$.



- This estimator requires no training.
- If we narrow the kernel with increasing amounts of data, the approach is consistent, i.e., will converge to some statistically optimal solution.

Summary

- Nadaraya-Watson kernel regression is an early precursor of the current attention mechanisms.
- It can be used directly with little to no training or tuning, both for classification and regression.
- The attention weight is assigned according to the similarity (or distance) between query and key.