Note from Miguel Carreira-Perpiñán

Nonparametric density estimation

- Given a sample $\mathcal{X} = \{\mathbf{x}_n\}_{n=1}^N$ drawn iid from an unknown density, we want to construct an estimator $p(\mathbf{x})$ of the density.
- *Histogram* (consider first $x \in \mathbb{R}$): split the real line into bins $[x_0 + mh, x_0 + (m+1)h]$ of width h for $m \in \mathbb{Z}$, and count points in each bin:

$$p(x) = \frac{1}{Nh}\,(\text{number of } x_n \text{ in the same bin as } x) \qquad x \in \mathbb{R}.$$

  - We need to select the bin width h and the origin x_0.
  - x_0 has a small but annoying effect on the histogram (near bin boundaries).
  - h controls the histogram smoothness: spiky if $h \downarrow$ and smooth if $h \uparrow$.
  - $p(x)$ is discontinuous at bin boundaries.
  - We don't have to retain the training set once we have computed the counts.
  - They generalize to D dimensions, but are practically useful only for $D \lesssim 2$
    In D dimensions, it requires an exponential number of bins, most of which are empty.
- *Kernel density estimate (Parzen windows)*: generalization of histograms to define smooth, multivariate density estimates. Place a kernel $K(\cdot)$ on each data point and sum them:

$$p(\mathbf{x}) = \frac{1}{Nh^D}\sum_{n=1}^N K\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right) \qquad \mathbf{x} \in \mathbb{R}^D \qquad \text{“sum of bumps”}.$$

  - K must satisfy $K(\mathbf{x}) \geq 0\ \forall \mathbf{x} \in \mathbb{R}^D$ and $\int_{\mathbb{R}} K(\mathbf{x})\,d\mathbf{x} = 1$. Typic. K is Gaussian or uniform.
    Gaussian: $K\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right) = (2\pi)^{-D/2}\exp\left(-\frac{1}{2}\|(\mathbf{x}-\mathbf{x}_n)/h\|^2\right)$. The uniform kernel gives a histogram without an origin x_0.
  - Only parameter: the bandwidth $h > 0$. The KDE is spiky if $h \downarrow$, smooth if $h \uparrow$.
    The KDE is not very sensitive to the choice of K.
  - $p(\mathbf{x})$ is continuous and differentiable if K is continuous and differentiable.
  - In practice, can take $K((\mathbf{x}-\mathbf{x}_n)/h) = 0$ if $\|\mathbf{x}-\mathbf{x}_n\| > 3h$ to simplify the calculation.
    We still need to find the samples $\mathbf{x}_n$ that satisfy $\|\mathbf{x}-\mathbf{x}_n\| \leq 3h$ (neighbors at distance $\leq 3h$).
  - Also possible to define a different bandwidth h_n for each data point $\mathbf{x}_n$ (*adaptive KDE*).
  - The KDE quality degrades as the dimension D increases (no matter how h is chosen).
    Could be improved by using a full covariance $\mathbf{\Sigma}_n$ per point, but it is preferable to use a mixture with $K < N$ components.
- *k-nearest-neighbor density estimate*: $p(\mathbf{x}) = \frac{k}{2N}\frac{1}{d_k(\mathbf{x})}$ for $\mathbf{x} \in \mathbb{R}^D$, where $d_k(\mathbf{x})$ = (Euclidean) distance of $\mathbf{x}$ to its kth nearest sample in $\mathcal{X}$.
  - Like using a KDE with an adaptive bandwidth $h = 2d_k(\mathbf{x})$.
    Instead of fixing h and counting how many samples fall in the bin, we fix k and compute the bin size containing k samples.
  - Only parameter: the number of nearest neighbors $k \geq 1$.
  - $p(\mathbf{x})$ has a discontinuous derivative. It does not integrate to 1 so it is not a pdf.