Learning with Density Matrices and Random Fourier Features

Fabio González

Machine Learning 2021-1

① Kernel density estimation (KDE)

Density estimation: Given a sample $\{x_i\}_{i=1..N}$ from an unknown distribution estimate the PDF of the distribution.

KDE (Rosenblat, 1956) (Parzen, 1962)

$$\hat{f}_X(x) = \frac{1}{N\lambda}\sum_{i=1}^{N} k_\lambda(x, x_i),$$

kernel Function

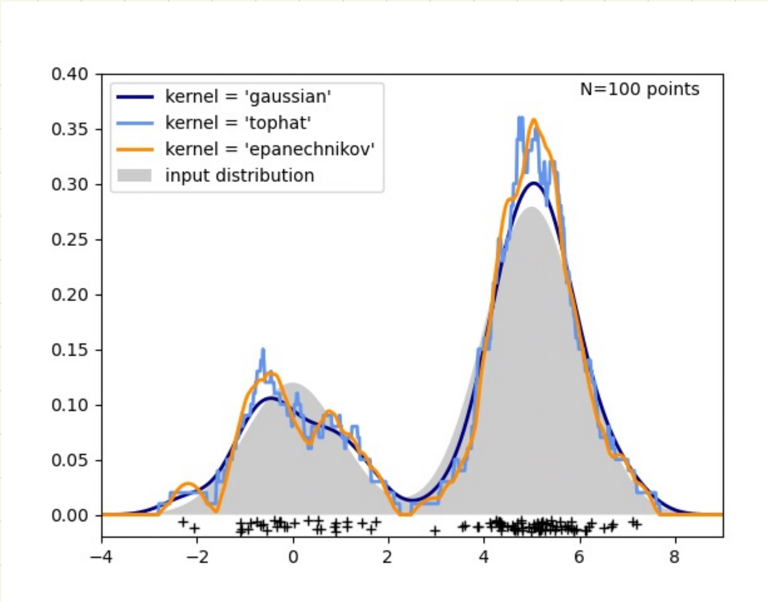
Bandwith

PDF estimator

$$\hat{g}_{\gamma,X}(x) = \frac{1}{N(\pi/\gamma)^{\frac{d}{2}}}\sum_{i=1}^{N} e^{-\gamma\|x_i - x\|^2}$$

Gaussian kernel

$\gamma = \frac{1}{2\sigma^2}$



Drawbacks

- Memory based method: you have to store all the training dataset.
- Prediction time $O(N)$
- Problems dealing with high-dimensional data

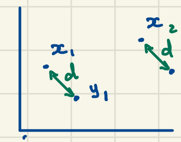
② Random Fourier Features (RFF) (Rahimi & Recht, 2007)

Theorem 1 (Bochner [13]). *A continuous kernel $k(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} - \mathbf{y})$ on $\mathcal{R}^d$ is positive definite if and only if $k(\delta)$ is the Fourier transform of a non-negative measure.*

→ Isotropic kernel

If a shift-invariant kernel $k(\delta)$ is properly scaled, Bochner's theorem guarantees that its Fourier transform $p(\omega)$ is a proper probability distribution. Defining $\zeta_\omega(\mathbf{x}) = e^{j\omega'\mathbf{x}}$, we have

$$k(\mathbf{x} - \mathbf{y}) = \int_{\mathcal{R}^d} p(\omega) e^{j\omega'(\mathbf{x}-\mathbf{y})}\, d\omega = E_\omega[\zeta_\omega(\mathbf{x})\zeta_\omega(\mathbf{y})^*], \tag{2}$$



$k(x_1, y_1) = k(x_2, y_2)$
$= k(d)$

if K is the Gaussian kernel, $K(x,y) = \langle \phi(x), \phi(y) \rangle_F$, the dimension of F is infinite

RFF method: Finds an embedding $\phi_{rff} : \mathbb{R}^d \to \mathbb{R}^D$ such that

$$\forall x, y \in \mathbb{R}^d \quad \langle \phi_{rff}(x), \phi_{rff}(y) \rangle \approx k(x,y)$$

$$\phi_{\text{rff}} : \mathbb{R}^d \to \mathbb{R}^D$$
$$x \mapsto \sqrt{\frac{2}{D}}(\cos(w_1^* x + b_1), \ldots, \cos(w_D^* x + b_D)) . \tag{1}$$

$w_1 \ldots w_D \sim P(w)$

$b_1 \ldots b_D \sim U(0, 2\pi)$

Advantage: kernel methods complexity is typically $> O(N^2)$

with RFF you can reduce this complexity

③ RFF and KDE

$$k^2_{\gamma/2}(x_i, x) = \left(e^{\gamma/2 \|x_i - x\|}\right)^2$$
$$= e^{\gamma \|x_i - x\|}$$
$$= k_\gamma(x_i, x)$$

$$\hat{g}_{\gamma, X}(x) = \frac{1}{M} \sum_{i=1}^{N} k_\gamma(x_i, x)$$
$$\approx \frac{1}{M} \sum_{i=1}^{N} \langle \phi_{rff}(x_i), \phi_{rff}(x) \rangle$$
$$\approx \langle \frac{1}{M} \sum_{i=1}^{N} \phi_{rff}(x_i), \phi_{rff}(x) \rangle$$

$$g_{\gamma, X}(x) = \frac{1}{M} \sum_{i=1}^{N} k^2_{\gamma/2}(x_i, x)$$
$$= \frac{1}{M} \sum_{i=1}^{N} \left(\phi_{rff}(x_i)^T \phi_{rff}(x) \right)^2$$
$$= \frac{1}{M} \sum_{i=1}^{N} \phi_{rff}(x)^T \underline{\phi_{rff}(x_i) \phi_{rff}(x_i)^T} \phi_{rff}(x)$$
$$= \phi_{rff}(x)^T \underbrace{\left[\frac{1}{M} \sum_{i=1}^{N} \phi_{rff}(x_i) \phi_{rff}(x_i)^T \right]}_{\rho} \phi_{rff}(x)$$

4 Density Matrices

The state of a quantum system is represented by a vector $\psi \in H$ (H is a Hilbert space, typically $\mathbb{C}^n$)

E.g. the spin of an electron $\{\uparrow, \downarrow\}$

$$\psi = (\alpha, \beta) \qquad |\alpha|^2 + |\beta|^2 = 1$$

Superposition: In general the a quantum state is a combination of basis states

$\uparrow : (1, 0) \qquad \downarrow : (0, 1) \qquad \psi = \alpha \uparrow + \beta \downarrow$

$|\alpha|^2$: Probability of obtaining $\uparrow$ $\qquad |\beta|^2$: Probability of obtaining $\downarrow$

Density Matrix: Representation of the state of a quantum system that can represent quantum uncertainty (superposition) and classical uncertainty.

$$\rho = \psi \psi^* = \begin{bmatrix} |\alpha|^2 & \alpha \beta^* \\ \beta^* \alpha & |\beta|^2 \end{bmatrix} \qquad \rho = \sum_{i=1}^{N} p_i \psi_i \psi_i^* \qquad \sum_{i=1}^{N} p_i = 1$$

Pure $\qquad\qquad$ mixed

Two systems: $\psi_1 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$ $\quad \rho_1 = \psi_1 \psi_1^* = \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix}$

$\psi_2 = (1, 0) \quad \psi_2' = (0, 1) \qquad \rho_2 = \frac{1}{2}\psi_2 \psi_2^* + \frac{1}{2}\psi_2' \psi_2'^* = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}$

Measurement probability (Born Rule):

$P(\varphi | \rho)$: Given a system in state ρ the probability of measuring φ

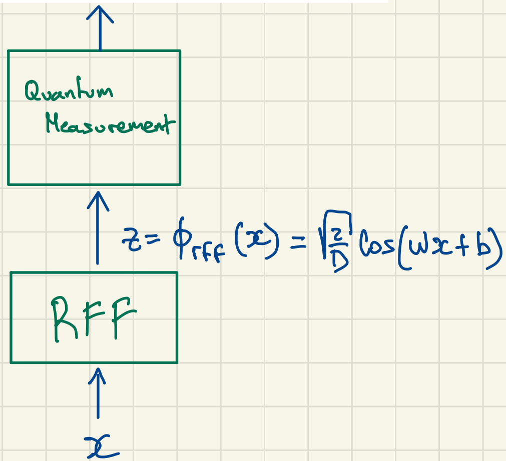
$$P(\varphi | \rho) = \mathrm{Tr}(\rho \varphi \varphi^*) = \varphi^* \rho \varphi$$

$\varphi = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) \qquad P(\varphi | \rho_1) = 1 \qquad P(\varphi | \rho_2) = \frac{1}{2}$

5 Density Matrix Kernel Density Estimation (DMKDE)

Prediction

$$\hat{f}_\rho(x) = \frac{\mathrm{Tr}(\rho\phi_{\mathrm{rff}}(x)\phi_{\mathrm{rff}}(x)^*)}{\mathcal{Z}} = \frac{\phi_{\mathrm{rff}}(x)^*\rho\phi_{\mathrm{rff}}(x)}{\mathcal{Z}}, \quad (12)$$



Training

- Input. A sample set $X = \{x_i\}_{i=1\ldots N}$ with $x_i \in \mathbb{R}^d$, parameters $\gamma \in \mathbb{R}^+$, $D \in \mathbb{N}$
- Calculate $W_{\mathrm{rff}} = [w_1 \ldots w_D]$ and $b_{\mathrm{rff}} = [b_1 \ldots b_D]$ using the random Fourier features method described in Section 2.1 for approximating a Gaussian kernel with parameters γ and D.
- Apply ϕ_{rff} (eq. (1)):

$$z_i = \phi_{\mathrm{rff}}(x_i). \quad (10)$$

- Density matrix estimation:

$$\rho = \frac{1}{N}\sum_{i=1}^{N} z_i z_i^*, \quad (11)$$

Time Complexity

Parzen's Estimator (KDE): $O(dN)$

DMKDE : $O(D^2)$

6 Factorized DMKDE

Spectral Decomposition:

Hermitian Matrix $\rightarrow \rho = V^T \Lambda V$ Eigenvalues Diagonal Matrix

Eigenvectors

$V \in \mathbb{R}^{r \times D}$ $\quad \Lambda \in \mathbb{R}^{r \times r}$ $\quad r < D$ Reduced dimension

$$\hat{f}_\rho(x) = \frac{1}{\mathcal{Z}} \|\Lambda^{\frac{1}{2}} V \phi_{\text{rff}}(x)\|^2$$

Time Complexity

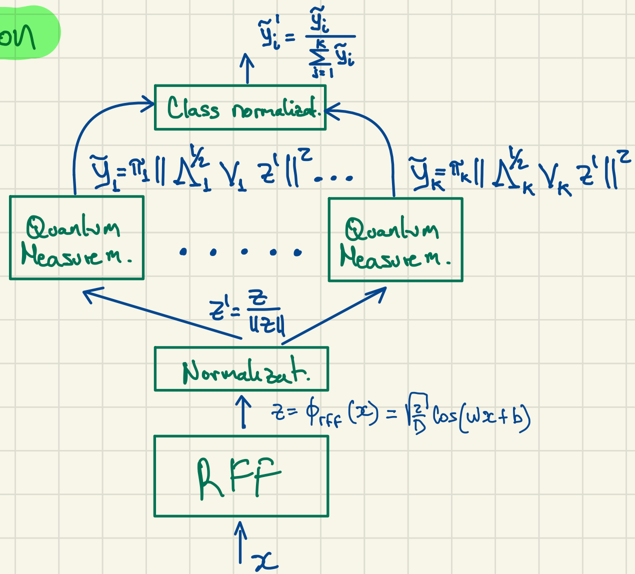
$O(Dr)$

8 Density Matrix Kernel Density Classification (DMKDC)

Kernel density classification

$$\underbrace{\hat{\Pr}(Y=j|X=x)}_{\text{Posterior probability}} = \frac{\pi_j \overbrace{\hat{f}_j(x)}^{\text{Density estimation}}}{\sum_{k=1}^{K} \underbrace{\pi_k}_{\text{Prior}} \hat{f}_k(x)}$$

Prediction



Training

Density Matrix Estimation

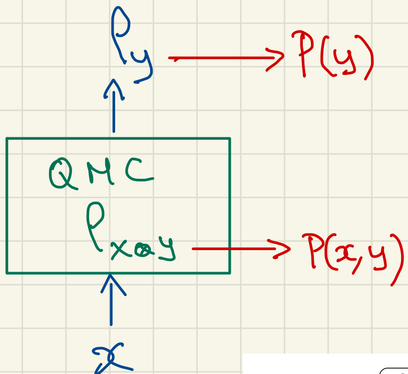
1. Use the RFF method to calculate W_{rff} and b_{rff}.
2. For each class i:
   (a) Estimate π_i as the relative frequency of the class i in the dataset.
   (b) Estimate ρ_i using eq. (11) and the training samples from class i.
   (c) Find a factorization of rank r of ρ_i:

$$\rho_i = V_i^* \Lambda V_i.$$

Stochastic Gradient Descent

$$L = \sum_{i=1}^{K} y_i \log(\tilde{y}_i)$$

9 Quantum measurement classification (QMC)



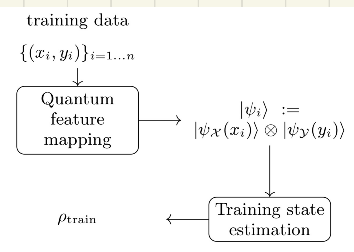


Figure 1. Training

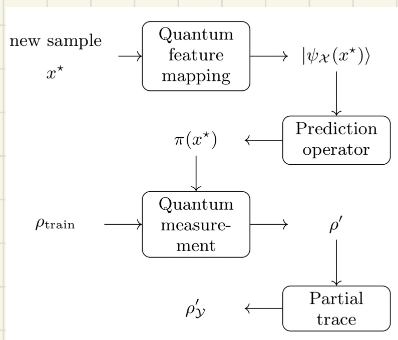


Figure 2. Prediction

- Generalizes DMKDC
- Produces a density matrix as output

Generalizes Bayesian Inference

Proposition 1. *Let $T = \{(x_i, y_i)\}_{i=1,\dots,n}$ be a set of training samples, $x^\star$ a sample to classify, with $x_i, x^\star \in \{1, \dots m\}$ and $y_i \in \{1, 2\}$. Let ρ_{train} be the state calculated using the mixed state, eq. (8) or equivalently the classic mixture eq. (9), and a one-hot encoding feature map for both x_i and y_i. Then the diagonal elements of the density matrix $\rho'_{\mathcal{Y}}$ calculated using eq. (12) correspond to an estimation, using Bayesian inference, of the conditional probabilities $P(y = i|x^\star)$:*

$$\rho'_{\mathcal{Y}i,i} = P(y = i|x^\star) = \frac{P(x^\star|y = i)P(y = i)}{P(x^\star)}, \qquad (13)$$

where $P(x^\star|y = i)$, $P(y = i)$ and $P(x^\star)$ are estimated from T.

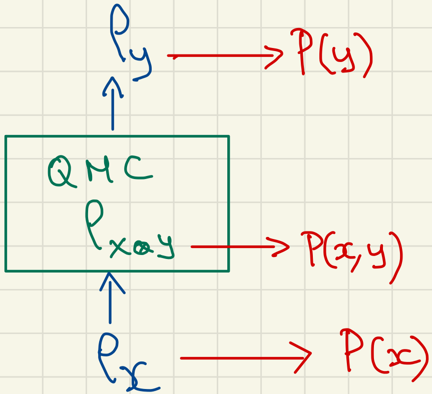
Can be seen as a kernel method

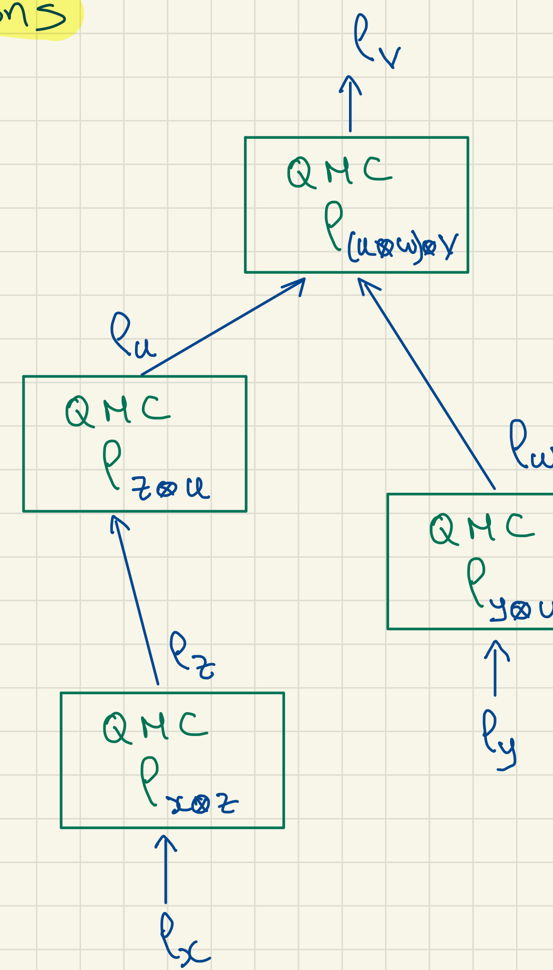
Proposition 2. *Let $T = \{(x_i, y_i)\}$ be a set of training samples, $x^\star$ a sample to classify, with $x_i, x^\star \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. Let ρ_{train} be the state calculated using a mixed state (eq. (8)) and quantum feature maps $\psi_{\mathcal{X}}$ and $\psi_{\mathcal{Y}}$. Then the density matrix $\rho'_{\mathcal{Y}}$, calculated with eq. (12), can be expressed as:*

$$\rho'_{\mathcal{Y}} = \mathcal{M} \sum_{i=1}^{N} |k(x^\star, x_i)|^2 \, |\psi_{\mathcal{Y}}(y_i)\rangle \langle \psi_{\mathcal{Y}}(y_i)| , \qquad (14)$$

where $k(x^\star, x_i) = \langle \psi_{\mathcal{X}}(x^\star)|\psi_{\mathcal{X}}(x_i)\rangle$ and $\mathcal{M}^{-1} = \mathrm{Tr}[\pi(x^\star)\rho_{train}\pi(x^\star)]$.

10 Further Generalizations





Multilayer Model (Bayesian network)

References

González, F. A., Vargas-Calderón, V., & Vinck-Posada, H. (2021). Classification with Quantum Measurements. Journal of the Physical Society of Japan, 90(4), 044002. https://arxiv.org/pdf/2004.01227.pdf

González, F. A., Gallego, A., Toledo-Cortés, S., & Vargas-Calderón, V. (2021). Learning with Density Matrices and Random Features. arXiv preprint arXiv:2102.04394 https://arxiv.org/abs/2102.04394

https://github.com/fagonzalezo/qmc