Project

→ Report 5 pages including references

- Motivation – what is the "learning" problem, what is the context?

  Should be stated in precise mathematical terms

- "Learning problem" Data space
  function space
  → Kernel & associated RKHS

- Kernel method

- Interpretation of results
  e.g. How does choice/design of kernel influence results?

→ Presentation (exactly same guidelines in 20 minutes)

Bayesian view of learning

- $\min_{f \in \mathcal{F}} \sum_{x \in S} \ell(x, f) \times \frac{1}{m}$

$S = \{(x_i, y_i)\}_{i \in [m]}$

$m = |S|$

gave point estimates $f(x)$ one value

replace with uncertainties

- $P(f(x) \mid S)$

$x_1, .., x_n \rightarrow$ test points

posterior prob. dist.

$$= \underbrace{\frac{1}{P(S)}}_{\substack{\text{normalization} \\ \text{constant} \\ \text{(independent of } f)}} P(\underbrace{y_1, \dots, y_m}_{\text{training points}} \mid f) \, P(f)$$

$P(f)$: Prior on f

- $P(y_1, .., y_m \mid f)$: likelihood

model : $y_i = f^*(x_i) + \varepsilon$

ε: Gaussian noise typically

when prior is

- $P(f)$: Gaussian process

ε : Gaussian noise $\rightarrow$ recovered regression

$\ell 2$ regularization as MAP estimate

- MAP estimate : $\max_{f} \underbrace{P(\{f(x)\} \mid S)}_{\text{Posterior using Bayes rule}}$

Sampling: Generating samples from a probab. distribution which is partially specified

π (target dist)

- $\pi \propto$ prior $\times$ likelihood

posterior (known) (known)

e.g. Gaussian process regression

statistical physics

$$\pi(x) = \frac{e^{-\beta E(x)}}{Z}$$

$$Z = \int e^{-\beta E(x)} dx$$

$\int \pi(x)\, dx = 1$ $E(x)$: energy

$x =$ [position(1), momentum(1), ...

position(m), momentum(m)]

$x \in \mathbb{R}^{2m}$ m: no of particles

- "Dynamic" transport of particles for high-dimensional sampling

$\pi(x) \propto e^{-\beta E(x)}$ β : inverse temperature

Score function : $\nabla \log \pi(x) = \nabla E(x) \times -\beta$

$$\pi(x) = \frac{e^{-\beta E(x)}}{Z}$$

$$\nabla \log \pi(x) = -\beta \nabla E(x) - \cancel{\log Z}$$

When $E(x)$ is known, but normalization Z is unknown, $\Rightarrow$ ~~the~~ score function is known

Score : Hyvarinen 2005

Score-matching

Generative Modeling: Want to sample π given $x_1, \dots, x_m \sim \pi$

MMD GANs

Ultimately, sampling is optimization

π : target

μ : reference
easy to sample.

Gaussian distribution e.g.

Transport maps: An invertible function $T: \mathbb{R}^d \circlearrowleft$
s.t. $T_{\#}\mu = \pi$. $\#$: pushforward

$\mu: \mathbb{R}^d \to \mathbb{R}^+$ $\pi: \mathbb{R}^d \to \mathbb{R}^+$

$x \in \mathbb{R}^d$ $x \sim \mu$

Then, $T(x) \sim \pi$

- $y \sim \pi$ $T^{-1}(y) \sim \mu$

$T(x) = (m_2 - m_1) + x$

$x \sim \mathcal{N}(m_1, 1)$ what is the density of $T(x)$?

$T(x) \sim \mathcal{N}(m_2, 1)$

$$\mathbb{E}T(x) = \mathbb{E}x + (m_2 - m_1)$$
$$= m_1 + m_2 - m_1 = m_2.$$

$$\mathrm{Var}(T(x)) = \mathrm{Var}\, x$$

$T(x) = ax + b$

$\mathrm{Var}\, T(x) = a^2 \mathrm{Var}(x)$

In general, what is T?

$$\min_{T \in \mathcal{T}} d(T_{\#}\mu, \pi)$$

if μ & π are densities,

e.g. KL divergence

$$d(\mu, \pi) = \int \log \frac{\mu}{\pi} \, \mu(x)\, dx$$

Integral prob. metrics.

$$d(\mu, \pi) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{x \sim \mu} f(x) - \mathbb{E}_{x \sim \pi} f(x) \right|$$

$\mathcal{F}$: Lip(1) d called Wasserstein-1 distance

$$\mathcal{F} = \{ f : \| f(x) - f(y) \| \le \| x - y \| \; \forall x, y \in \mathbb{R}^d \}$$

Remark: with an optimal T, evaluate it on samples from μ.

$x \sim \mu$ $T(x) \sim \pi$.

$\min_{h \in \mathcal{H}} \frac{1}{\lvert S \rvert} \sum_{x \in S} \ell(x, h)$	$\min_{T \in \mathcal{T}} d(T_{\#}\mu, \pi)$
h : soln of above optimization	T : soln of above optimization
$\mathbb{E}_{x \sim \mathcal{D}} \ell(x, h)$: generalization, how good h is.	Estimate $d(T_{\#}\mu, \pi)$ on samples!

Estimators for distances/metrics on spaces of probability densities

- Empirical density:

$$x_1, \dots, x_m \sim \pi$$

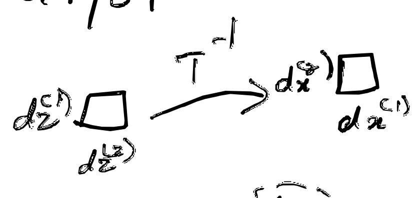
$$\hat{\pi}_m = \mathrm{Unif}(x_1, \dots, x_m)$$

$$\mathop{\mathbb{E}}_{x \sim \pi} f(x) \approx \frac{1}{m} \sum_{i \leq m} f(x_i)$$

- Change of variables formula:

$$\underset{\substack{\uparrow \\ \text{prob. density}}}{T_\# \mu} = \underset{\substack{\uparrow \\ \text{prob. density}}}{\pi}$$

$$\pi = \frac{\mu \circ T^{-1}}{|\det dT| \circ T^{-1}}$$



$$\mathop{\mathbb{E}}_{x \sim \pi} f(x) = \int f(x)\, \pi(x)\, dx$$

$$z = T^{-1}(x)$$

$$= \int f \circ T(z)\, \pi(T(z)) |\det dT(z)|\, dz$$

$$f(T(z))$$

$$d(T_\# \mu, \pi)$$

$$= d\left(\frac{\mu \circ T^{-1}}{|\det dT| \circ T^{-1}}, \pi \right)$$

Can use KL divergence,
W^1 distance

different distances give rise to different optimization problems.

SVGD

$$d(\mu, \pi) = \max_{\phi \in \mathcal{F}} \mathop{\mathbb{E}}_{x \sim \mu} (\mathrm{tr}\, A_\pi \phi(x))$$

$$A_\pi \phi(x) = \nabla_x \log \pi(x)\, \phi(x)^T + \nabla_x \phi(x)$$

Stein operator

$$\mathop{\mathbb{E}}_{x \sim \pi} A_\pi \phi(x) - \int \nabla \log \pi(x)\, \phi(x)^T \pi(x)\, dx + \int \nabla_x \phi(x)\, \pi(x)\, dx$$

How to solve optimization problem?

$$\min_{h \in \mathcal{H}} \sum_{x \in S} \ell(x, h) \times \frac{1}{m}$$

To solve:
parametrize $\mathcal{H}$.

linear reg

$$\mathcal{H} = \{ \underline{\underline{w}}^T x + b : w \in \mathbb{R}^d, b \in \mathbb{R} \}$$

kernel reg

$$\mathcal{H} = \{ \underline{\underline{w}}^T \phi(x) : w \in \mathbb{R}^D \}$$

and solve optimization problem for w

$$\min_{T \in \mathcal{T}} d(T_\# \mu, \pi)$$

parameterize $\mathcal{T}$