MLPR (Non-Bayesian)

Linear Regression: $f = X\underline{w} + b$ or $f(\underline{x}) = \sum_k w_k \phi_k(\underline{x})$

RBF: $\phi(\underline{x}) = \exp(-(\underline{x}-\underline{c})^T(\underline{x}-\underline{c})/h^2)$

Sigmoid: $\sigma(\underline{x}) = 1/(1+\exp(-\underline{v}^T\underline{x} - b))$

L2: $\hat{E}(\underline{w}) = E(\underline{w}) + \lambda \underline{w}^T\underline{w}$; or $y' = \begin{bmatrix} y \\ 0_K \end{bmatrix}$; $\phi' = \begin{bmatrix} \phi \\ \sqrt{\lambda}\, \mathbb{I}_K \end{bmatrix}$

Bias: for μ, var. scales w/ $1/N$

Gaussian: $p(x) \propto e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$; $p(\underline{x}) \propto e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T\Sigma^{-1}(\underline{x}-\underline{\mu})}$

+ve definite: $\underline{z}^T \Sigma \underline{z} \geq 0 \;\; \forall\, \underline{z} \in \mathbb{R} \neq 0$

1-hot: $y = [0\ 0 \ldots 1 \ldots 0]^T$

Naive Bayes:

→ binary $p(\underline{x} \mid y=k, \theta) = \prod_d \theta_{d,k}^{x_d}(1-\theta_{d,k})^{1-x_d}$

→ cond. indep. given class

Gaussian Generative: $p(\underline{x} \mid y=k) = \mathcal{N}(\underline{x}, \underline{\mu}^{(k)}, \Sigma^{(k)})$

$\Rightarrow p(y=k \mid \underline{x}) \propto p(\underline{x} \mid y=k)\, \pi_k \to \pi_k \approx \frac{\sum_n \mathbb{I}(y^{(n)}=k)}{N}$

SGD vs. **Analytic** (fight)

$\underline{w}^* = (X^T X)^{-1} X^T \underline{y}$

$\underline{w}^{(t+1)} \leftarrow \underline{w}^{(t)} - \eta \nabla_{\underline{w}}[\underline{r}^T\underline{r}] \to \underline{r} = \underline{y} - X\underline{w}$

Logistic: $f(\underline{x}; \underline{w}) = \sigma(\underline{w}^T\underline{x})$

$L(\underline{w}) = p(\underline{y} \mid X, \underline{w})$; $LL(\underline{w}) = -\sum_n \log \sigma(z^{(n)} \underline{x}^{(n)} \underline{w}^T)$

$z^{(n)} = 2y^{(n)} - 1$

Softmax: $f_k = e^{\underline{x}^T\underline{w}^{(k)}} / \sum_{k'} e^{\underline{x}^T\underline{w}^{(k')}}$

Robust

$p(y=+1 \mid \underline{x}, \underline{w}, m) = \begin{cases} \sigma(\underline{w}^T\underline{x}) & m=1 \\ \frac{1}{2} & m=0 \end{cases}$; $p(m) = \begin{cases} 1-\epsilon & m=1 \\ \epsilon & m=0 \end{cases}$

Convex: $C(\alpha \underline{w} + (1-\alpha)\underline{w}') \le \alpha C(\underline{w}) + (1-\alpha) C(\underline{w}')$

Neural Nets

$$h^{(\ell)} = g^{(\ell)}\left(W^{(\ell)} \underline{h}^{(\ell-1)} + \underline{b}^{(\ell)}\right) ; \quad h^{(0)} = \underline{x}, \quad \ell = 1 \dots N$$

Backprop

FW mode

$\frac{\partial \cdot}{\partial u} \to$ scalar $\to \forall$

$$\boxed{\frac{\partial f}{\partial u_i} = \sum_{d=1}^{D} \frac{\partial f}{\partial x_d} \frac{\partial x_d}{\partial u_i}}$$

BW mode (Reverse)

$\frac{\partial z}{\partial \cdot} \to$ error w.r.t. $\forall$

$$C = AB, \Rightarrow \bar{A} = \bar{C} B^T, \quad \bar{B} = A^T \bar{C}$$

$$C = g(A) \Rightarrow \bar{A} = g'(A) \odot \bar{C}$$

- $\underset{L \times M}{A} \times \underset{M \times N}{B}$ is $O(LMN)$

Autoencoders

NN, hidden layer $K \ll D$ for dim reduction

L1 $\to$ sparse, if $K = 2$, visualize

PCA

$\Sigma = \frac{1}{N} X^T X$; $V = \text{eig}(\Sigma)$; project & profit

center first

$VV^T = \mathbb{I}$

SVD

$X \approx U S V^T$

U	$N \times K$	eigenvectors of XX^T
S	$K \times K$	diagonal
V^T	$K \times D$	eigenvectors of $X^T X$

PCA is SVD : $X \approx X V V^T$

PPCA

$\to$ probabilistic

- assume K-dim Gaussian $\underline{\nu} \sim \mathcal{N}(\underline{0}, \mathbb{I}_K)$
- project onto D $\quad \underline{x} = W \underline{\nu}$ (W is $D \times K$)
- Then $\underline{x} \sim \mathcal{N}(\underline{0}, WW^T + \sigma^2 \mathbb{I})$ $\to$ add some noise (for a cheeky encore)

MLPR (Bayesian)

• Linear Regression

$$p(y|\underline{x}, \underline{w}) = \mathcal{N}(y; f(x; \underline{w}), \sigma_y^2)$$ → pinch of noise

1) w/ MLE ≡ regular LR

2) + prior $p(\underline{w}) = \pi(\underline{w}) = \mathcal{N}(\underline{w}; \underline{0}, \sigma_w^2 \mathbb{I})$

$$p(\underline{w}|D) \propto p(D|\underline{w})\, p(\underline{w})$$ → Bayesian conspiracy

b/p of Gaussian $\left\{ \propto \mathcal{N}(\underline{w}; \underline{0}, \sigma_w^2 \mathbb{I})\, \mathcal{N}(y; \Phi \underline{w}, \sigma_y^2 \mathbb{I}) \right.$ → likelihood

w/ w_N, V_N

3) ~~Inference~~ Predictions over posterior: $P(y|D) = \int p(y|\underline{x}, \underline{w})\, p(\underline{w}|D)\, d\underline{w}$ (test, test, posterior)

4 (optional)) Model choice over prior : $p(y|x, M) = \int p(y|\underline{x}, \underline{w}, M)\, p(\underline{w}|M)\, d\underline{w}$ (train, prior)

$|||$ $P(D|M)$

5) Cost functions:

$\hat{y}$ is a point estimate; optimize $\mathbb{E}_{p(y|D)}[L(y, \hat{y})]$

6) Cheeky-breeky (if linear)

$$p(y|D, \underline{x}) = \mathcal{N}(y; \underline{x}^T \underline{w}_N, \underline{x}^T V_N \underline{x} + \sigma_y^2)$$

7) for linear $(p(D|M))^{-1} = \dfrac{P(\underline{w}|D)}{p(\underline{w})\, p(D|\underline{w})}$ → triple Gauss

Gaussian Processes

1) function $\underline{f}$ as vector, $\operatorname{cov}[f_i, f_j] = k(\underline{x}^{(i)}, \underline{x}^{(j)})$

2) $f \sim GP$ iff $p(\underline{f}) = \mathcal{N}(\underline{f}; \underline{0}, K)$

3) Pinch a noise + test points

$$p\left(\begin{bmatrix} y \\ \underline{f}^* \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} y \\ \underline{f}^* \end{bmatrix}; \underline{0}, \begin{bmatrix} K(X,X) + \sigma_y^2 & K(X, X^*) \\ K(X^*, X) & K(X^*, X^*) \end{bmatrix}\right)$$

(noise; test; smudge)

Kernels

linear : $k(\underline{x}^{(i)}, \underline{x}^{(j)}) = \sigma_w^2\, \phi(\underline{x}^{(i)})^T \phi(\underline{x}^{(j)}) + \sigma_b^2$

gaussian: $k(\underline{x}^{(i)}, \underline{x}^{(j)}) = \underbrace{\sigma_f^2}_{\text{function scale}} \exp\left(-\frac{1}{2}\sum_{d=1}^{D} (x_d^{(i)} - x_d^{(j)})^2 / \underbrace{\ell_d^2}_{\text{length btw. training points}}\right)$

$p(\underline{y} \mid X, \theta) \rightarrow$ marginal LH, hyperparameters

Logistic Regression

MAP: $\underline{w}^* = \arg\max_{\underline{w}} [\log p(\underline{w} \mid D)] = \ldots$ Bayes

MLE: no prior

approximated by

$\boxed{p(\underline{w} \mid D)} \propto \prod \sigma(\underline{w}^T \underline{x}^{(n)}) \, \mathcal{N}(\underline{w}; \underline{0}, \sigma_w^2 \mathbb{I})$

to find

Laplace Approx.

1) Define $E(\underline{w}) = -\log \underbrace{p(\underline{w}, D)}_{\propto\, p(D)\, p(\underline{w})}$ $\nearrow$ same as $p(\underline{w} \mid D)$ up to a const.

2) $\underline{w}^* = \arg\min_{\underline{w}} E(\underline{w})$

3) $H_{ij} = \left. \frac{\partial E(\underline{w})}{\partial w_i \partial w_j} \right|_{\underline{w} = \underline{w}^*}$

4) $p(\underline{w} \mid D) \approx \mathcal{N}(\underline{w}; \underline{w}^*, H^{-1})$

$p(\underline{w} \mid D) = \frac{p(\underline{w}, D)}{p(D)}$

At $\underline{w} = \underline{w}^*$

$p(D) \approx \frac{p(\underline{w}, D)}{p(\underline{w} \mid D)}$

$\approx \frac{p(\underline{w}^*, D)\, (2\pi)^{D/2}}{|H|^{1/2}}$

5) predictions

$p(y{=}1 \mid \underline{x}, D) \approx \int \sigma(\underline{w}^T \underline{x}) \, \mathcal{N}(\underline{w}; \underline{m}, V) \, d\underline{w}$

$= \mathbb{E}_{\mathcal{N}(\underline{w}; \underline{m}, V)} [\sigma(\underline{w}^T \underline{x})]$

$= \int \sigma(a) \, \mathcal{N}(a; \underline{m}^T \underline{x}, \underline{x}^T V \underline{x}) \, da$

approximates $p(\underline{w} \mid D)$

Variational Inference

1) Define $q(\underline{w}; \alpha = \{\underline{m}, V\}) = \mathcal{N}(\underline{w}; \underline{m}, V)$

$D_{KL}(p \| q) = \int p(\underline{z}) \log \frac{p(\underline{z})}{q(\underline{z})} \, dz \geq 0$ (Gibbs)

if $p(\underline{w} \mid D)$, intractable

2) Minimize

$D_{KL}(q(\underline{w}; \alpha) \| p(\underline{w} \mid D)) = -\int d\underline{w} \, \underbrace{q(\underline{w}; \alpha) \log p(\underline{w} \mid D)}_{q \text{ is high where } p \neq 0} + \underbrace{\int d\underline{w} \, q(\underline{w}; \alpha) \log q(\underline{w}; \alpha)}_{\text{high entropy, spread out}}$

Variational Inference (ctnd.)

$$p(\underline{w}|D) = \frac{p(D|\underline{w})\,p(\underline{w})}{p(D)}$$

$$D_{KL}(q\|p) = \underbrace{\mathbb{E}_q[\log q(\underline{w})] - \mathbb{E}_q[\log p(D|\underline{w})] - \mathbb{E}_q[\log p(\underline{w})]}_{J(q)}$$

$$+ \log p(D) \rightarrow \text{const. w.r.t. } q$$

- Bound on $p(D)$ / marginal LH

$$D_{KL}(q\|p) \geq 0, \quad \Rightarrow \log p(D) \geq \underbrace{-J(q)}$$

ELBO: evidence lower bound

- Hardest term is $\mathbb{E}_q[\log p(D|\underline{w})] =$

$$= \sum_{n=1}^{N} \mathbb{E}_q[\log P(y^{(n)} \,|\, \underline{x}^{(n)}, \underline{w})]$$

- Minimize $J(\underline{m}, V)$ w.r.t. $\underline{m}, V$
  - $V = LL^T, \quad L = \begin{cases} \log(L_{ii}) & i = j \\ L_{ij} & i \neq j \end{cases}$
- Reparametrization trick for

$$\mathbb{E}_{\mathcal{N}(\underline{w};\,\underline{m},V)}[f(\underline{w})] = \mathbb{E}_{\mathcal{N}(\underline{\nu};\,0,\mathbb{I})}[f(\underline{m} + L\underline{\nu})]$$

sample $\underline{w}$, by $\nu = \mathcal{N}(0,\mathbb{I})$

$$\underline{w} = \underline{m} + L\underline{\nu}$$

- Monte-Carlo Estimate

$$\ldots \approx \frac{1}{S}\sum_{s=1}^{S} f(\underline{m} + L\underline{\nu}^{(s)}), \quad \underline{\nu}^{(s)} \sim \mathcal{N}(0,\mathbb{I})$$

$$\approx f(\underline{m} + L\underline{\nu})$$ • single unbiased estimate

- Compute $\nabla_{\underline{m}}[\ldots]$, $\nabla_L[\ldots]$ & SGD

Gaussian Mixture Models

$z^{(n)} \in \{1, \dots, K\} \rightarrow$ classes

- MLE: $p(\underline{x}^{(n)} | \theta) = \sum_k \pi_k \mathcal{N}(z^{(n)}; \mu^{(k)}, \Sigma^{(k)})$

Expectation Maximization

$r_k^{(n)} = \delta_{z^{(n)}, k} \rightarrow$ responsibility for cluster k

$\pi_k = \frac{r_k}{N}$, $r_k = \sum_{n=1}^{N} r_k^{(n)}$, $\mu^{(k)} = \frac{1}{r_k} \sum_{n=1}^{N} r_k^{(n)} \underline{x}^{(n)}$

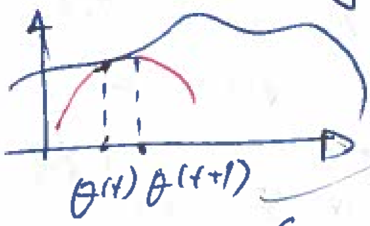
$$\Sigma^{(k)} = \frac{1}{r_k} \sum_{k=1}^{N} r_k^{(n)} \underline{x}^{(n)} \underline{x}^{(n)T} - \mu^{(k)} \mu^{(k)T}$$

- E-step: $r_k^{(n)} = p(z^{(n)} = k | \underline{x}^{(n)}, \theta) = \dfrac{\pi_k \mathcal{N}(\underline{x}^{(n)}; \mu^{(k)}, \Sigma^{(k)})}{\sum_l \pi_l \mathcal{N}(\underline{x}^{(n)}; \mu^{(l)}, \Sigma^{(l)})}$
- M-step: update

L1: $c(\underline{w}) = E(\underline{w}) + \lambda \|\underline{w}\|_1$

Newton: $\underline{w}^{(t+1)} \leftarrow \underline{w}^{(t)} - H^{-1} f$

Bound-based:



$\rightarrow$ needs tight bound

Ensembling: $p(y | \underline{x}, \mathcal{D}) \approx \frac{1}{S} \sum_{s=1}^{S} p(y | \underline{x}, \underline{w}^{(s)})$, $\underline{w}^{(s)} \sim p(\underline{w} | \mathcal{D})$

Bagging:
- sample w/ replacement from N points S times
- fit S models, then poolfit

Mixture of experts: $p(y | \underline{x}, \theta) = \sum_k p(y | \underline{x}, z=k, \theta) \overbrace{p(z=k | \underline{x}, \theta)}^{\text{N.N.}}$

Boosting: $F_{m+1}(x) = F_m(x) + h(x) \rightarrow$ weak model

fit $F_{m+1}(x)$ to $y - F_m(x) = h(x)$