1 Maximum Likelihood Estimator

Three properties of maximum likelihood esti-

- 1. θ_{ML} is consistent. $\theta_{ML} \to \theta_0$ when $n \to \infty$.
- 2. θ_{ML} is asymptotically normal. $\sqrt{n}(\theta_{ML} \theta_0$) ~ $\mathcal{N}(0, I_n(\theta_0))$ when $n \to \infty$ and $I_n(\theta_0)$ is the fisher information.
- 3. θ_{ML} is asymptotically efficient. θ_{ML} mi- 3 BLR and GP nimizes $\mathbb{E}(\theta - \theta_0)^2$ when $n \to \infty$ because the asymptotic variance equals the Rao-Cramer bound (MLE is asymptotically unbiased). Note: when *n* is finite, θ_{ML} is not necessarily efficient, e.g., Stein estimator is universally more efficient for single sam-

Rao-Cramer bound: for any unbiased estimator $\hat{\theta}$ of θ_0 , $\mathbb{E}(\hat{\theta} - \theta_0)^2 \geq 1/I_n(\theta_0)$, where $I_n(\theta) = -\mathbb{E}(\frac{\partial^2}{\partial \theta^2} \log f(X; \theta) \mid \theta) =$ $\mathbb{E}(\frac{\partial}{\partial \theta} \log f(X; \theta) \mid \theta)^2$ is the fisher informati-

Sketch of Proof: define $\Lambda = \frac{\partial \log P(X;\theta)}{\partial \theta}$. Cauchy-Schwarz says $Cov^2(\Lambda, \hat{\theta}) \leq Var(\Lambda)Var(\hat{\theta}) =$ $\mathbb{E}(\Lambda^2)$ **Var** $(\hat{\theta})$ because $\mathbb{E}\Lambda = 0$. Note that $\mathbf{Cov}(\Lambda, \hat{\theta}) = \mathbb{E}(\Lambda \hat{\theta}) = \int_{X} \hat{\theta}(x) \frac{\partial}{\partial \theta} f(x; \theta) dx =$ $\frac{\partial}{\partial \theta} \int_X \hat{\theta}(x) f(x;\theta) dx = \frac{\partial}{\partial \theta} \mathbb{E} \hat{\theta} = 1$. Therefore, A function is a kernel iff (1) symmetry $\operatorname{Var}(\hat{\theta}) \geq 1/\mathbb{E}(\Lambda^2)$.

However, when the dimension of problem goes to infinity while keeping the data-dim ratio fixed, MLE is biased and the *p*-values are unreliable.

2 Regression

Bias-Variance trade-off

Let D be the training dataset and \hat{f} be the predictive function. $\mathbb{E}_D \mathbb{E}_{Y|X}(\hat{f}(X) - Y)^2 =$ $\mathbb{E}_D \mathbb{E}_{Y|X} [(\hat{f}(X) - \mathbb{E}_{Y|X}Y)^2 + (\mathbb{E}_{Y|X}Y - Y)^2] =$ $\mathbb{E}_D(\hat{f}(X) - \mathbb{E}(Y \mid X))^2 + \mathbb{E}_D(\mathbb{E}(Y \mid X) - Y)^2 =$ $\mathbb{E}_D(\hat{f}(x) - \mathbb{E}_D\hat{f}(x))^2 + \left(\mathbb{E}_D\hat{f}(x) - \mathbb{E}(Y \mid X)\right)^2 +$ $\mathbb{E}_D(\mathbb{E}(Y \mid X) - Y)^2$. It means that expected square error (training) = variance of prediction + squared bias + variance of noise.

The optimal trade-off is achieved by avoiding under-fitting (large bias) and over-fitting (large variance). Note that here the variance of output is computed by refitting the regressor on a new dataset.

Regularization

Ridge and Lasso can be viewed as MAP (maximum a posterior) estimation. A Gaussian prior on β is equivalent to Ridge and a Laplacian prior is equivalent to Lasso. Using

SVD, we get Ridge has built-in model selection: $X\beta^{\text{Ridge}} = \sum_{j=1}^{d} [d_j^2/(d_j^2 + \lambda)] u_j u_j^T Y$ (each $u_i u_i^T Y$ can be viewed as a model). Lasso has more sparse estimations because the gradient of regularization does not shrink as in the case of Ridge.

Bayesian Linear Regression

Model $Y = X\beta + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Prior $\beta \sim$ $\mathcal{N}(0,\Lambda^{-1})$. Posterior $\beta \mid X,Y \sim \mathcal{N}(\mu_{\beta},\Sigma_{\beta})$, where $\mu_{\beta} = (X^T X + \sigma^2 \Lambda)^{-1} X^T Y$ and $\Sigma_{\beta} =$ $(\sigma^{-2}X^TX + \Lambda)^{-1}$.

Gaussian Process

 $Y = \begin{pmatrix} Y_0 \\ Y_n \end{pmatrix}$ is the combination of observed and prediction value. Assume a Gaussian prior of $\mathcal{N}(0, K + \sigma^2 I)$, where $K_{ij} = k(x_i, x_j)$ is kernel. GP regression is the conditional/Posterior distribution on Y_0 , $\mathbb{E}[Y_1|Y_0] = K_{10}(\sigma^2 I_0 +$ $(K_{00})^{-1}Y_0$, $Cov[Y_1] = \sigma^2I_1 + K_{11} - K_{10}(\sigma^2I_0 + I_0)$ $(K_{00})^{-1}K_{01}$. Bayesian LR is a special case of GP with linear kernel $k(x, y) = x^{\top} \Lambda^{-1} y$.

Kernel Function

k(x, x') = k(x', x) and (2) semi-positive definite $\int_{\Omega} k(x,x')f(x)f(x')dxdx' \ge 0$ for any $f \in L_2$ and $\Omega \in \mathbb{R}^d$ (continuous) or $K(X) \geq 0$ (discrete). The latter is equivalent to (1) $a^{\top}Ka \geq$ $0, \forall a \text{ or } (2) k(x, x') = \phi(x)^T \phi(x') \text{ for some } \phi.$

Kernel Construction

If $k_{1,2}$ are valid kernels, then followings are valid: (1) $k(x, x') = k_1(x, x') + k_2(x, x')$. (2) k(x, x') = $k_1(x,x') \cdot k_2(x,x')$. Proof: expand by Mercer's thm. (3) $k(x, x') = ck_1(x, x')$ for constant c > 0. (4) $k(x,x') = f(k_1(x,x'))$ if f is a polynomial with positive coefficients or the exp. Proof: polynomial can be proved by applying the product, positive scaling and addition. Exp can be proved by taking limit on the polynomial. (5) $k(x,x') = f(x)k_1(x,x')f(x')$. (6) $k(x, x') = k_1(\phi(x), \phi(x'))$ for any function ϕ .

Example: RBF kernel $k(x,y) = e^{-\|x-y\|^2/2\sigma^2} =$ $e^{-\|x\|^2/2\sigma^2} \times e^{x^T y/2\sigma^2} \times e^{-\|y\|^2/2\sigma^2}$ is valid. (1) $x^T y$ linear kernel is valid (2) then $\exp(\frac{1}{\sigma^2}x^Ty)$ is valid, (3) let $f(x) = \exp(-\frac{1}{2\sigma^2}||x||^2)$, by rules f(x)k(x,y)f(y) RBF is valid.

Mercer's Theorem: Assume k(x, x') is a valid kernel. Then there exists an orthogonal basis e_i and $\lambda_i \geq 0$, s.t. $k(x, x') = \sum_i \lambda_i e_i(x) e_i(x')$.

Linear Methods for Classification Concept Comparison

- 1. Probabilistic Generative, modeling p(x,y): (1) can create new samples, (2) outlier detection, (3) probability for prediction, (4) high computational cost and (5) high bias.
- 2. Probabilistic Discriminative, modeling $p(y \mid x)$: (1) probability for prediction, (2) medium computational cost and (3) medium bias.
- 3. Discriminative, modeling y = f(x): (1) no probability for prediction, (2) low computational cost and (3) low bias.

Infer p(x, y) for classification problems

Use $p(x,y) = p(y)p(x \mid y)$. Since y has finite states, model p(y) and p(x | y) for different y. The modeling requires to (1) guess a distribution family and (2) infer parameters by

Compute $p(y \mid x)$ by discriminant analysis (DA) Linear DA

Goal: classify a sample into two Gaussian distribution with $\Sigma_0 = \Sigma_1$. After calculation, $p(y = 1 \mid x) = 1/(1 + \exp(-\log \frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0)})) =$

 $1/(1 + \exp(w_1^T x + w_0))$ since the quadratic term is eliminated due to $\Sigma_0 = \Sigma_1$.

Ouadratic DA

Goal: classify a sample into two Gaussian distribution with $\Sigma_0 \neq \Sigma_1$. After calculation, $p(y = 1 \mid x) = 1/(1 + \exp(x^T W x + w_1^T x + w_0)).$

Optimization Methods

Optimal Learning Rate for Gradient Descent

Goal: find $\eta^* = \operatorname{argmin}_{\eta} L(w^k - \eta \cdot \nabla L(w^k))$.

By Taylor expansion of $L(w^{k+1})$ at w^k and solve for the optimal η , we get $\eta^* =$

$$\frac{\|\nabla L(w^k)\|^2}{\nabla L(w^k)^T H_L(w^k) \nabla L(w^k)}.$$

However, naive gradient descent has two weaknesses: (1) it often has a zig-zag behavior, especially in a very narrow, long and slightly downward valley; (2) the gradient update is small near the stationary point. This can be mitigated by adding a momentum term in the update: $w^{k+1} = w^k - \eta \nabla L(w^k) + \mu^k (w^k - w^{k-1})$ which speeds the update towards the "common" direction.

Newton's Method

Taylor-expand L(w) at w_k to derive the opti- **Problem** $\min_x f(x)$ s.t. $g_{i \in [I]}(x) \leq 0$ and mal w^{k+1} : $L(w) \approx L(w^k) + (w - w^k)^T \nabla L(w^k) +$ $\frac{1}{2}(w-w^k)^T H_I(w^k)(w-w^k) \implies w^{k+1} = w^k - 1$

 $H_I^{-1}(w^k)\nabla L(w^k)$.

Pros: (1) better updates compared to GD since it uses the second Taylor term and (2) does not require learning rate.

Cons: requires H_I^{-1} which is expensive.

Bavesian Method

In most cases, the posterior is intractable. Use approximation of posterior instead.

Laplacian Method

Idea: approximate posterior near the MAP estimation with a Gaussian distribution. $p(w \mid$ $(X,Y) \propto p(w,X,Y) \propto \exp(-R(w))$, where R(w) = $-\log p(w, X, Y)$. Let $w^* = \operatorname{argmin} R(w)$ be the MAP estimation and Taylor-expand R(w) at w^* : $R(w) \approx R(w^*) + \frac{1}{2}(w - w^*)^T H_R(w^*)(w - w^*)$. Therefore, $p(w \mid X, Y) \propto \exp(-R(w^*) - \frac{1}{2}(w - w^*))$ $(w^*)^T H_R(w^*)(w-w^*)$ and thus $(w \mid X, Y) \sim$ $\mathcal{N}(w^*, H_R^{-1}(w^*)).$ AIC & BIC

- Define BIC = $k \log N 2 \log \hat{L}$, where k is #parameters and \hat{L} is the likelihood $p(x \mid w^*)$. A lower BIC means a better model.
- Define AIC = $2k 2\log \hat{L}$. A lower AIC means a better model.

LDA by loss minimization

Perceptron

Goal: for $y_i \in \{0,1\}$, find w, s.t. $y_i w^T x_i > 0$ for any i. The classification function is c(x) = $sgn(w^Tx)$.

L(y,c(x)) = 0 if $yw^Tx > 0$ and L(y,c(x)) = $-yw^Tx$ o.w. By gradient descent, the Perceptron is guaranteed to converge if (1) the data is linearly separable, (2) learning rate $\eta(k) > 0$, (3) $\sum_{k} \eta(k) \rightarrow +\infty$ and (4) $(\sum_k \eta(k)^2)/(\sum_k \eta(k))^2 \rightarrow 0$. However, there exists multiple solutions if the data is linearly separable.

Fisher's LDA

Idea: project the two distribution into one dimension and maximize the ratio of the variance between the classes and the variance within the classes, i.e., $\max(w^T u_1 - w^T u_0)^2 / (w^T S w)$, where $S = \Sigma_0 + \Sigma_1$. Let gradient be zero and solve for w^* , we get $w^* \propto S^{-1}(u_1 - u_0)$.

We first compute w^* and fit distributions of the two-class projection. Then apply Bayesian decision theory to make classification.

5 Optimization with Constraint

 $h_{i \in [I]}(x) = 0$. Solve it with **KKT Cond**: (1) Stationary $\nabla f + \sum_{i} \lambda_{i} \nabla g_{i} + \sum_{i} \mu_{i} \nabla h_{i} = 0$, (2)

 $h_i(x) = 0$, (3) primal feasibility $g_i(x) \le 0$, (4) dual feasibility $\lambda_i \geq 0$, (5) complementary slackness $\lambda_i g_i(x) = 0$.

Weak Duality: Lagrangian $L(x, \lambda, \mu) = f(x) + \frac{1}{2} f(x)$ $\lambda^{\top} g(x) + \mu^{\top} h(x), \lambda > 0$. Dual function $F(\lambda, \mu) :=$ $\min_{x} L(x, \lambda, \mu)$. Denote \tilde{x} optima of original problem, then $\lambda^{\top} g(\tilde{x}) + \mu^{\top} h(\tilde{x}) \leq 0, \forall \lambda, \mu$, $F(\lambda, \mu) = \min_{x} L(x, \lambda, \mu) \le L(\tilde{x}, \lambda, \mu) \le f(\tilde{x}) =$ $\min_{x,h(x)=0,g(x)<0} f(x)$

Strong Duality in Convex Optimization

If **Slater's cond** (1) f convex (2) g convex (3) h linear (4) $\exists \overline{x}$ s.t. $g_i(\overline{x}) < 0$ and $h_i(\overline{\mathbf{x}}) = 0$, then Strong Duality $\max_{\lambda,\mu} F(\lambda,\mu) = 0$ $\min_{x,h(x)=0,g(x)\leq 0} f(x)$ holds.

6 Support Vector Machine **Linear Separable Case**

Primal: $\max_{w,b} \left\{ \frac{1}{\|w\|} \min_i y_i(w^\top x_i + b) \right\} \Leftrightarrow$ $\max_{w,b,t} t \text{ s.t. } \forall i, t \leq y_i(w^\top x_i + b) \text{ and } ||w|| = 1$ $\Leftrightarrow \min_{w,h} \frac{1}{2} w^2 \text{ s.t. } \forall i, 1 \leq y_i (w^\top x_i + b)$ (1) KKT cond: $\forall i, \alpha_i \geq 0, (1 - y_i(w^{\top}x_i + b)) \leq$ $0, \alpha_i (1 - y_i (w^{\top} x_i + b)) = 0$ (2) **Dual**: $\max_{\alpha} \sum_{i} \alpha_{i} - \frac{1}{2} \sum_{i,j} \alpha_{i} \alpha_{j} y_{i} y_{j} K(x_{i}, x_{j})$ s.t. $(\alpha_i \ge 0) \land (\sum_i \alpha_i y_i = 0)$

Non-separable Case

Introduce slack variables $\xi_i := \max\{1 - 1\}$ $y_i(w^{\top}x_i + b), 0\} = [1 - y_i(w^{\top}x_i + b)]_+ \text{ into loss.}$ **Primal:** $\min_{w,h} \frac{1}{2} w^2 + C \sum_i \xi_i = \min_{w,h} \frac{1}{2} w^2 +$ $C[1-y_i(w^{\top}x_i+\bar{b})]_+$. Hinge loss $[1-x]_+$. Equivalent form: $\min_{w,h} \frac{1}{2}w^2 + C\sum_i \xi_i$ s.t. $y_i(w^\top x_i + b) \ge 1 - \xi_i$ and $\xi_i \ge 0$

Dual: $\max_{\alpha} \sum_{i} \alpha_{i} - \frac{1}{2} \sum_{i,j} \alpha_{i} \alpha_{j} y_{i} y_{j} K(x_{i}, x_{j})$ s.t. $\sum_i \alpha_i y_i = 0$ and $0 \le \alpha_i \le C$

Multi-class SVM

 $\min_{w=[w_{0:K-1}],b=[b_{0:K-1}]} \frac{1}{2} ||w||^2 + \sum_i C\xi_i \text{ s.t. } \xi_i \ge 0$ and $(w_{v_i}^\top x + b_{v_i}) - (w_v^\top x + b_v) \ge 1 - \xi_i, \forall y \ne y_i$

Structural SVM

y is structured, e.g. trees, maximum margin between y_i, y_i depends on their similarity, so the condition changes to $w^{\top}\Psi(x_i,y_i)$ – $w^{\top}\Psi(x_i, y) \geq \Delta(y_i, y) - \xi_i, \ \forall \ y \neq y_i.$

7 Ensemble

Bagging Each bagged estimator have bias $\beta = \mathbb{E}(y - b(x))^2$, variance $\sigma^2 = \text{Var}b(x)$ covariance $\rho^2 = \text{Cov}(b(x), b'(x))/\sigma^2$. Then $\mathbb{E}(y - y)$ $\sum_{m} b^{(m)}(x)/M)^2 = \beta^2 + \sum_{m} \mathbb{E}(\beta - b^{(m)}(x))^2/M^2 =$ $\beta^2 + \sigma^2/M + \sigma^2 \rho^2 (1 - 1/M)$. In class we assume $\rho = 0$. Anyway Bagging reduces variance.

Random Forest is a case of Bagging. Bagging induces implicit regularization.

 $f_t(x) = \operatorname{argmin}_{b(x)} \sum w_i^{(t)} \mathbb{I}_{\{v_i \neq b(\mathbf{x}_i)\}}$ (2) error $\epsilon_t =$ $(\sum w_i^{(t)} \mathbb{I}_{\{v_i \neq f_t(x_i)\}}) / \sum w_i^{(t)}$ (3) estimator weight $\alpha_t =$ $\log(\frac{1-\epsilon_t}{\epsilon_t})$ (4) data weight $w_i^{(t+1)} = w_i^{(t)} e^{\alpha_t \mathbb{I}_{\{y_i \neq f_t(\mathbf{x}_i)\}}}$ **Prediction** $\hat{c} = \operatorname{sgn}(\sum_{t=1}^{M} \alpha_t f_t(\mathbf{x}))$

Gradient Boosting Initial $f_0(x) = 0$. For $t \in [M]$, (1) train $(\alpha_t, b^{(t)}) \leftarrow \operatorname{arg\,min}_{\alpha > 0, h \in \mathcal{H}}$ $\sum_{i=1}^{n} L(y_i, \alpha b(x_i) + f_{t-1}(x_i))$ (2) update function $f_t(x) \leftarrow \alpha_t b^{(t)}(x) + f_{t-1}(x)$. Prediction $\hat{c}(x) =$ $\operatorname{sgn}(f_M(x))$. Adaboost is GB with $L(y, \hat{y}) = e^{-y\hat{y}}$.

8 Generative Models

ELBO $\log p(y) = \log \int p(y \mid \theta) p(\theta) d\theta =$ $\log \mathbb{E}_{\theta \sim q} \left[p(y \mid \theta) \frac{p(\theta)}{a(\theta)} \right] \geq \mathbb{E}_{\theta \sim q} \left[\log \left(p(y \mid \theta) \frac{p(\theta)}{a(\theta)} \right) \right]$ $\mathbb{E}_{\theta \sim q}[\log p(y \mid \theta)] - KL(q || p(\cdot))$

VAE Goal: Find a latent representation z of *x* with simple prior $p_{\theta}(z)$. Problem: $p_{\theta}(x) =$ $\mathbb{E}_{\theta} p(x|z)$ intractable. Solution: use encoder net $q_{e}(x|z)$ and $q_{d}(z|x)$ to model conditional and posterior prob.

ELBO for VAE training loss $l = \sum \log(p_{\theta}(x_i))$

$$\begin{split} &\log\left(p_{\theta}\left(x_{i}\right)\right) = \mathbb{E}_{Z \sim q_{\phi}\left(z|x_{i}\right)}\left[\log p_{\theta}\left(x_{i}\right)\right] = \mathbb{E}_{Z}\left[\log\frac{p_{\theta}\left(x_{i}\mid z\right)p_{\theta}(z)}{p_{\theta}\left(z\mid x_{i}\right)}\right] \\ &= \mathbb{E}_{Z}\left[\log\frac{p_{\theta}\left(x_{i}\mid z\right)p_{\theta}(z)}{p_{\theta}\left(z\mid x_{i}\right)}\frac{q_{\phi}\left(z\mid x_{i}\right)}{q_{\phi}\left(z\mid x_{i}\right)}\right] \\ &= \mathbb{E}_{Z}\left[\log p_{\theta}\left(x_{i}\mid z\right)\right] - \mathbb{E}_{Z}\left[\log\frac{q_{\phi}\left(z\mid x_{i}\right)}{p_{\theta}(z)}\right] + \mathbb{E}_{Z}\left[\log\frac{q_{\phi}\left(z\mid x_{i}\right)}{p_{\theta}\left(z\mid x_{i}\right)}\right] = \\ &= \mathbb{E}_{Z}\left[\log p_{\theta}\left(x_{i}\mid z\right)\right] - D_{KL}\left(q_{\phi}\left(z\mid x_{i}\right)\|p_{\theta}(z)\right) + D_{KL}\left(q_{\phi}\left(z\mid x_{i}\right)\|p_{\theta}\left(z\mid x_{i}\right)\right) \\ &= \mathcal{L}\left(x_{i}, \theta, \phi\right) &\geq 0 \end{split}$$

Generative Adversarial Network: Generator G and Discriminator D. Optimize $\min_G \max_D V(D,G)$ where V(D,G) = $\mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]$

9 Convergence of SGD, Robbins-Monro

Loss gradient $\ell(\cdot)$, SGD update $z^{(t)} \leftarrow \ell\theta^{(t)} +$ $\gamma^{(t)}, \theta^{(t+1)} \leftarrow \theta^{(t)} - \eta(t)z^{(t)}, \gamma^{(t)}$ noise.

Problem: Whether $\theta^{\infty} \to \arg_{\theta^*} \mathbb{E}[\ell(\theta^*)] \triangleq 0$? Assume: (1) $\mathbb{E}[\gamma] = 0$, (2) $\mathbb{E}[\gamma^2] = \sigma$ (3) $(\theta - \theta^*)\ell(\theta) > 0, \forall \theta \neq \theta^*$ (4) $\exists b, \ell(\theta) < b, \forall \theta$. If $(1) \eta^{(t)} \to 0 \ (2) \ \sum \eta(t) = \infty \ (3) \ \sum \eta^2(t) < \infty,$

then $\mathbb{P}\left(\theta^* = \theta^{(t)}\right) \underset{t \to \infty}{\longrightarrow} 1$.

Proof: $\mathbb{E}[(\theta^{(t+1)} - \theta^*)^2] = \mathbb{E}[((\theta^{(t)} - \theta^*) - \eta(t)l(\theta^{(t)}) - \theta^*)]$ $\eta(t)\gamma^{(t)}$. $\gamma^{(t)}$ independent with $\theta^{(t)}$, $\ell(\theta^{(t)})$, **Adaboost** Initial $w_t^{(0)} = 1/n$. For $t \in [M]$, (1) train so LHS = $\mathbb{E}[(\theta^* - \theta^{(t)})^2] - 2\eta(t)\mathbb{E}[\ell(\theta^{(t)})(\theta^* - \theta^{(t)})^2]$ $|\theta^{(t)}| + \eta^2(t) (\mathbb{E}[\ell^2(\theta^{(t)})] + \mathbb{E}[\gamma^2(t)]) \le \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 1$ $2\sum_{i \le t} \eta(i) \mathbb{E}[\ell(\theta^{(i)})(\theta^* - \theta^{(i)})] + \sum_{i \le t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \le \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] < -\infty, 0 = \lim \mathbb{E}[\ell(\theta^{(i)})(\theta^* - \theta^{(i)})] =$ $\lim \mathbb{P}(\theta^* = \theta^{(i)}) \mathbb{E}[\ell(\theta^{(i)})(\theta^* - \theta^{(i)})|\theta^* = \theta^{(i)}] + \mathbb{P}(\theta^* \neq \theta^{(i)})$ $\theta^{(i)})\mathbb{E}[\ell(\theta^{(i)})(\theta^* - \theta^{(i)})|\theta^* \neq \theta^{(i)}], \lim \mathbb{P}(\theta^* \neq \theta^{(i)}) = 0$

10 Non-parametric Bayesian Inference (BI) **Exact Conjugate Prior of Multivariate Gaussian** Data: $x_i \sim \mathcal{N}(\mu, \Sigma)$ i.i.d.. Inverse Wishart: $\Sigma \sim$ $W^{-1}(S, v) \propto |\Sigma|^{(v+p+1)/2} \exp(-\text{Tr}(\Sigma^{-1}S)/2).$

Normal Inverse Wishart as conjugate prior: $p(\mu, \Sigma | m_0, k_0, v_0, S_0) = \mathcal{N}(\mu | m, \Sigma / k_0) \mathcal{W}^{-1}(\Sigma | S_0, v_0)$ Condition on θ , Margin over G: θ_{n+1} | Update rule: $m_p = (k_0 m_0 + N \bar{x})/(k_0 + N), k_p =$ $k_0 + N$, $v_p = v_0 + N$, $S_p = S_0 + k_0 m_0 m_0^{\top} - k_p m_p m_p^{\top} +$ $\sum (x_i - \overline{x})(x_i - \overline{x})^{\top}$.

BI with Semi-Conjugate Prior

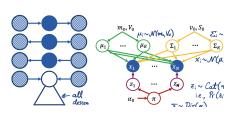
New prior: $\mu \sim \mathcal{N}(m_0, V_0)$, $\Sigma \sim \mathcal{W}^{-1}(S_0, v_0)$, then posterior $p(\mu, \Sigma | X)$ is intractable, but condition posterior is exact, $p(\mu|\Sigma, X) = \mathcal{N}(m_p, V_p)$, $V_p^{-1} = V_0^{-1} + N\Sigma^{-1}, V_p^{-1}m_p = V_0^{-1}m_0 + N\Sigma^{-1}\overline{x};$ $p(\Sigma|\mu,X) = W^{-1}(S_p,v_p), v_p = v_0 + N, S_p = S_0 +$ $\sum x_i x_i^{\top} + N \mu \mu^{\top} - 2N \overline{x} \mu^{\top}$.

Gibbs sampling: random variable $p(z_1, \dots, z_n)$ intractable, cyclically resample z_i according to tractable conditional distribution $p(z_i|z_{i})$ n times, when $n \to \infty$, $(z_1, \dots, z_p) \sim p(z_1, \dots, z_p)$ Finally, replace posterior with MC sampling: $\mathbb{E}_{\theta|X} f(x|\theta) \approx \sum f(x|\theta_i)/N$

BI for Gaussian Mixture Model

Data model: latent K class variable $z_i \sim$ Cat(π), observed $x_i \sim \mathcal{N}(\mu_{z_i}, \Sigma_{z_i})$. Prior: $\mu_k \sim$ $\mathcal{N}(m_0, V_0), \ \Sigma_k \sim \mathcal{W}^{-1}(S_0, v_0), \ \pi \sim \operatorname{Dir}(\alpha) \propto$ $\prod_{k=1}^{K} p_{k}^{\alpha_{k}-1}$. Prior also intractable.

Goal Gibbs sampling for BI, but to simplify conditional distribution.



d-seperation: for verifying conditional independence. Given with observed variable set C, if every path from variable A to B is blocked $\mathbf{P}(\mathcal{R}(\hat{c}_n^*) - \inf_{c \in \mathcal{C}} \mathcal{R}(c) > \epsilon) \le 9n^d \exp\left(-\frac{n\epsilon^2}{32}\right)$

on probability graph, then A and B are independent condition on C. By this thm: (1) z_i , z_i (2) μ , π (3) Σ , π all independent condition on other parameter. Sampling procedure: $(1) z^{(t)} \leftarrow p(\cdot | x, \mu^{(t-1)}, \Sigma^{(t-1)}), (2) \mu^{(t)} \leftarrow p(\cdot | x, \Sigma^{(t-1)}, z^{(t)}),$ $(3) \Sigma^{(t)} \leftarrow p\left(\cdot|x,\mu^{(t)},z^{(t)}\right), (4) \pi^{(t)} \leftarrow p\left(\cdot|x,z^{(t)}\right)$

BI for Non-Parametric GMM

Goal: sample from infinite categorical distri. Dirichlet Process (DP): Θ parameter space, H prior distri on Θ , A_1, \dots, A_r arbitrary partition of Θ . G a categorical distribution over $\{A_i\}$ is $G \sim \mathrm{DP}(\alpha, H)$ if $(G(A_1), \ldots, G(A_r)) \sim$ $Dir(\alpha H(A_1),...,\alpha H(A_r)).$

Posterior:
$$G|\{\theta_i\}_{i=1}^n \sim DP\left(\alpha + n, \frac{\alpha H + \sum_{i=1}^n \delta_{\theta_i}}{\alpha + n}\right)$$

 $\theta_1, \dots, \theta_n \sim \frac{1}{\alpha+n} \left(\alpha H + \sum_{i=1}^n \delta_{\theta_i} \right)$, Leads to CRP

Three Methods of Sampling from DP

In $K \to \infty$ GMM, θ in DP is z, G is π .

(1) Chinese Restaurant Process (CRP), sample z, marginalize over π :

$$p(z_n = k | \theta_{i < n}) = \begin{cases} n_k / (\alpha + n - 1), \text{ existing } k \\ \alpha / (\alpha + n - 1), \text{ new } k \end{cases}$$

Expect # of Class $\sum_{i=1}^{n} \frac{\alpha}{\alpha+i-1} \simeq \alpha \log \left(1+\frac{n}{\alpha}\right)$

(2) Stick-breaking Construction samples π : $\beta_k \sim \text{Beta}(1, \alpha), \ \theta_k^* \sim H, \ \pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l)$

(3) Marginalize over μ , Σ when sampling z (if intractable), less variance (Rao-Blackwall).

Exchangeability: $p(\{\theta_i\}) = \prod_{n=1}^N p(\theta_n | \{\theta_{i \le n}\})$ unchanged after permuting sampling order. DeFinetti's Thm any exchangeable distri is a mixture model $P(\{\theta_i\}) = \prod_{i=1}^n G(\theta_i) dP(G)$

11 PAC Learning

- A learning algorithm A can learn $c \in C$ if there is a poly(.,.), s.t. for (1) any distribution \mathcal{D} on $\bar{\mathcal{X}}$ and (2) $\forall \epsilon \in [0, 1/2], \delta \in [0, 1/2],$ \mathcal{A} outputs $\hat{c} \in \mathcal{H}$ given a sample of size at least poly($\frac{1}{c}$, $\frac{1}{\delta}$, size(c)) such that $P(\mathcal{R}(\hat{c}) \inf_{c \in C} \mathcal{R}(c) \leq \epsilon \geq 1 - \delta.$
- A is called an efficient PAC algorithm if it runs in polynomial of $\frac{1}{6}$ and $\frac{1}{\delta}$.
- \mathcal{C} is (efficiently) PAC-learnable from \mathcal{H} if there is an algorithm A that (efficiently) learns C from \mathcal{H} .
- Finite C, $\mathbf{P}(\mathcal{R}(\hat{c}_n^*) \inf_{c \in C} \mathcal{R}(c) > \epsilon) \leq$ $2|\mathcal{C}|\exp\left(-\frac{n\epsilon^2}{2}\right)$ is PAC-learnable.
- C with $\dim_{VC} = d < \infty$ is PAC-learnable,

A Appendix

$$\begin{array}{l} \text{(1) } \partial_x(AB) = A\partial_x B + (\partial_x A)B, \text{(2) } \partial_x A^{-1} = -A^{-1}(\partial_x A)A^{-1}, \\ \text{(3) } \partial_x \ln \det A = \operatorname{Tr} \left(A^{-1}\partial_x A\right), \end{array}$$

(3)
$$\partial_x \ln \det A = \operatorname{Tr} \left(A^{-1} \partial_x A \right)$$

$$\begin{aligned} & \text{Define } (\partial_A f)_{ij} := \partial_{a_{ji}} f \text{, then (4)} \ \partial_A \text{Tr}(BA) = \partial_A \text{Tr}(AB) = \\ & B \text{, (5)} \ \partial_A \ln \det A = A^{-1} \text{, (6)} \ \partial_A \text{Tr}(ABA^\top) = (B + B^\top) A^\top, \end{aligned}$$

$$\begin{split} \mathcal{N}(\mu, \Sigma) &= (2\pi)^{-d/2} |\Sigma|^{-1/2} e^{-(x-\mu)^T \Sigma^{-1} (x-\mu)/2}, \text{ Conditional } \\ \mathbb{E}[y_2|y_1] &= \mu_2 + \Sigma_{21} \Sigma_{11}^{-1} (y_1 - \mu_1), \text{ Cov}[y_2 \mid y_1] = \Sigma_{22} - \\ \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}. \text{ Marginal } \mathbb{E}(y_2) = \mu_2, \text{ Cov}[y_2] = \Sigma_{22} \end{split}$$

$$(A+UC^{-1}V)^{-1}=A^{-1}-A^{-1}U(C+VA^{-1}U)^{-1}VA^{-1}.$$