# AML Cheat Sheet

## Prob., distributions and identitites

| | |
|---|---|
| SVD | $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top, \mathbf{U} \in \mathbb{R}^{n\times d}, \mathbf{V} \in \mathbb{R}^{d\times d}$ |
| Cauchy-Schwarz | $\left\|\sum_i u_i \overline{v}_i\right\|^2 \le \sum_j \|u_j\|^2 \sum_k \|v_k\|^2$ |
| Cov. (univariate) | $Cov[x,y] = E[(x - E[x])(y - E[y])$ |
| Cov. (mult'vari.) | $Cov[x,y] = E_{x,y}[xy^\top] - E_x[x]E_y[y]^\top$ |
| | $V[x \pm y] = V[x] + V[y] \pm 2Cov[x,y]$ |
| | $V_x[Ax + b] = V_x[Ax] = AV_x[x]A^\top$ |
| Sum Rule | $P(X = x) = \sum_y p(X = x, Y = y)$ |
| Conditional | $P(X\|Y) = P(X,Y)/P(Y)$ |
| Bayes' Rule | $P(Y\|X) = \frac{P(X\|Y)P(Y)}{P(X)}$ |
| Multi Gaussian | |
| $p(\mathbf{x}\|\boldsymbol{\mu},\boldsymbol{\Sigma}) = ((2\pi)^d \cdot$ | $\|\boldsymbol{\Sigma}\|)^{-1/2} \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}))$ |
| $P(x_i\|c,\mu_c,\sigma_c^2) =$ | $\frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{(x_i-\mu_c)^2}{2\sigma_c^2}\right)$ |
| Markov | $P[X \ge \epsilon] \le \mathbb{E}[X]/\epsilon, X \ge 0, \epsilon > 0$ |
| Hoeffding L. | $\mathbb{E}[e^{sX}] \le \exp(s^2(b-a)^2/8)$ |
| Hoeffding Thm | |
| $P[S_n - \mathbb{E}S_n \ge t] \le \exp(-2t^2/\sum(b_i-a_i)^2)$ | |
| $P[S_n - \mathbb{E}S_n \le -t] \le \exp(-2t^2/\sum(b_i-a_i)^2)$ | |
| if $S_n \to \bar{X}_n$ | $t \to n\epsilon$ |

## Kernels

- Gaussian (RBF) kernel: $k(x,x') = \exp(-\|x-x'\|_2^2/h^2)$
- Dimension of $k(\mathbf{x},\mathbf{x}') = (\mathbf{x}^\top\mathbf{x}' + c)^d$ is $\binom{N+d}{d}$

## Properties

| | |
|---|---|
| Symmetry | $k(x,x') = k(x',x)$ |
| Pos semi-def | $\int_\Omega k(\mathbf{x}',\mathbf{x}) f(\mathbf{x}) f(\mathbf{x}') d\mathbf{x} d\mathbf{x}', f \in L_2, \Omega \subseteq \mathbb{R}^d$ |
| construct | $\theta : \mathbf{x}_i \mapsto (\sqrt{\lambda_t}\mathbf{v}_{it})_{t=1}^n$ |

## Identities

| | |
|---|---|
| Addition | $k(x,x') = k_1(x,x') + k_2(x,x')$ |
| Multiply | $k(x,x') = k_1(x,x')k_2(x,x')$ |
| Scalar | $k(x,x') = ck_1(x,x')$ for $c > 0$ |
| Transform | $k(x,x') = f(k_1(x,x'))$ |
| | $f$ polynom with positive coeff. or exp. |
| func. multiply | $f(x)k_1(x,x')f(x')$ for any $f$ |

## Risks

$$Q(y,f(x)) = \begin{cases} (y-f(x))^2 & \text{quadratic loss (regr.)} \\ \mathbb{I}_{\{y\neq f(x)\}} & \text{0-1 loss (class.)} \\ \exp(-\beta y f(x)) & \text{exponential loss (class.)} \end{cases}$$

| | |
|---|---|
| Cond. Exp. Risk | $R(f,X) = \int_Y Q(Y,f(X))P(Y\|X)dY$ |
| (Total) Exp. Risk | $R(f) = \mathbb{E}_X[R(f,X)]$ |
| Emp. Error | $\hat{R}(f,X) = \frac{1}{n}\sum_{i=1}^n Q(y_i, f(X_i))$ |

## Maximum Likelihood Estimators

$\hat{\theta}^{ML} \in \arg\max P(\mathcal{X}|\theta) \overset{i.i.d.}{=} \prod_{i=1}^n P(x_i|\theta)$

Definitions:

| | |
|---|---|
| Bias | $bias(\hat{\theta}_n) = \mathbb{E}[\hat{\theta}_n] - \theta$ |
| | $bias[\hat{f}(x)] = \mathbb{E}_D\hat{f}(x) - \mathbb{E}[Y\|x]$ |
| Consistency | $\forall\epsilon,\ P[\|\hat{\theta}_n - \theta\| > \epsilon] \overset{n\to\infty}{\to} 0$ |
| Score | $\Lambda(\theta) := \frac{\partial \log P(x\|\theta)}{\partial\theta}$ |
| • $E_{X\|\theta}[\Lambda] = 0$; | • $E_{X\|\theta}[\Lambda] = \frac{\partial}{\partial\theta}b_{\hat{\theta}} + 1$ |
| Fisher Information | $I(\theta) = \mathbb{V}\left[\frac{\partial \log P(x\|\theta)}{\partial\theta}\right]$ |
| Asymptotical efficiency | $\lim_{n\to\infty}(\mathbb{V}[\hat{\theta}_n(x_1,...,x_n)]I(\theta))^{-1} = 1$ |

Results

| | |
|---|---|
| Rao-Cramer | $\mathbb{E}_{X\|\theta}[(\hat{\theta}-\theta)^2] \ge (1 + \frac{\partial}{\partial\theta}b_{\hat{\theta}})^2/I^{(n)}(\theta) + b_{\hat{\theta}}^2$ |
| ML converg. | $\sqrt{n}(\hat{\theta}_n^{ML} - \theta_0) \overset{D}{\to} \mathcal{N}(0, J^{-1}(\theta_0)I(\theta_0)J^{-1}(\theta_0))$ |
| | $J(\theta) = -\mathbb{E}[\frac{\partial^2 \log P(x\|\theta)}{\partial\theta\partial\theta^T}]$ |
| ML consist. | $\hat{\theta}_n^{ML} \overset{p}{\to} \theta_0$ |
| If $\hat{\theta}^{ML}$ for $\theta$ | $g(\hat{\theta}^{ML})$ for $g(\theta)$ |

## Bayesian Learning

| | |
|---|---|
| Maximum a Posteriori | $\hat{\theta} \in \arg\max_\theta p(x\|\theta)p(\theta)$ |
| Prediction: | |
| $p(X = x\|\mathcal{X}) =$ | $\int p(x\|\theta)p(\theta\|\mathcal{X})d\theta$ |
| Rec. Bayesian Est. | |
| $p(\theta\|\mathcal{X}^n) =$ | $\frac{p(x_n\|\theta)p(\theta\|\mathcal{X}^{n-1})}{\int p(x_n\|\theta)p(\theta\|\mathcal{X}^{n-1})d\theta}$ |

## Regression

| | |
|---|---|
| $\varepsilon \sim \mathcal{N}(0,\sigma)$ | $y = \mathbf{X}\beta + \varepsilon$ |
| Least-square fit | $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ |
| $f^* =$ opt. estimator | $\hat{\beta} \sim \mathcal{N}(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$ |
| For $\hat{\theta} = \mathbf{c}^T\mathbf{y}$ unbiased: | $\mathbb{V}[\mathbf{a}^T\hat{\beta}] \le \mathbb{V}[\mathbf{c}^T\mathbf{y}]$ |
| MSE | $\mathbb{E}_D\mathbb{E}_{X,Y}(\hat{f}(X) - Y)^2 =$ |
| variance | $+\mathbb{E}_X\mathbb{E}_D(\hat{f}(X) - \mathbb{E}_D\hat{f}(X))^2$ |
| $bias^2$ | $+\mathbb{E}_X(\mathbb{E}_D\hat{f}(X) - \mathbb{E}_Y[Y\|X])^2$ |
| noise | $+\mathbb{E}_{X,Y}(Y - \mathbb{E}_Y[Y\|X])^2$ |
| Ridge | $\hat{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$ |
| Gen. Reg. | $\hat{\beta} = \arg\min_\beta RSS(\beta) + \lambda\sum_{j=1}^d \|\beta_j\|^q$ |
| $bias(\hat{f}) = E[\hat{f}] - f^*$ | $Var[\hat{f}] = E[(\hat{f} - E[\hat{f}])^2$ |

## Bayesian Linear Regression

$$Model:\ Y = \mathbf{X}\beta + \epsilon,\ \epsilon \sim \mathcal{N}(0,\sigma^2)$$

$$Likelihood:\ p(Y|\mathbf{X},\beta,\sigma^2) = \mathcal{N}(\mathbf{X}\beta,\sigma^2)$$

$$Prior:\ p(\beta|\mathbf{\Lambda}) = \mathcal{N}(0,\mathbf{\Lambda}^{-1})$$

$$Posterior:\ p(\beta|\mathbf{X},\mathbf{y},\mathbf{\Lambda}) = \mathcal{N}(\mu_\beta,\Sigma_\beta)$$

$$\mu_\beta = (\mathbf{X}^T\mathbf{X} + \sigma^2\mathbf{\Lambda})^{-1}\mathbf{X}^T\mathbf{y}\Sigma_\beta = \sigma^2(\mathbf{X}^T\mathbf{X} + \sigma^2\mathbf{\Lambda})^{-1}$$

$$(\beta-\mu_\beta)^\top\Sigma_\beta^{-1}(\beta-\mu_\beta) = \beta^\top\Sigma_\beta^{-1}\beta - 2\beta^\top\Sigma_\beta^{-1}\mu_\beta + \mu_\beta^\top\Sigma_\beta^{-1}\mu_\beta$$

Conditioning a Gaussian: $p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\mu_{a|b},\Sigma_{a|b})$
$\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(\mathbf{x}_b - \mu_b), \Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$
$\Lambda_{aa} = (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}, \Lambda_{ab} = -\Lambda_{aa}\Sigma_{ab}\Sigma_{bb}^{-1}$

## Gaussian Process

Joint distribution of $[\mathbf{y}, y_{n+1}]$ is given by
$\mathbf{y}|\mathbf{X},\sigma^2 \sim \mathcal{N}(0, \mathbf{X}^T\Lambda^{-1}\mathbf{X} + \sigma^2\mathbf{I})$, kernelized version:

$$p\left(\begin{bmatrix}\mathbf{y}\\ y_{n+1}\end{bmatrix}|x_{n+1},\mathbf{X},\sigma^2\right) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}\mathbf{C}_n & \mathbf{k}\\ \mathbf{k}^T & c\end{bmatrix}\right)$$

$$\mathbf{C}_n = \mathbf{K} + \sigma^2\mathbf{I}_n \quad c = k(x_{n+1},x_{n+1}) + \sigma^2$$

$$\mathbf{k} = k(x_{n+1},\mathbf{X}) \quad \mathbf{K} = k(\mathbf{X},\mathbf{X})$$

$$p\left(\begin{bmatrix}\mathbf{a}_1\\ \mathbf{a}_2\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{a}_1\\ \mathbf{a}_2\end{bmatrix}|\begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{bmatrix}, \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\right)$$

$\mathbf{a}_1, \mathbf{u}_1 \in \mathbb{R}^e$; $\Sigma_{11} \in \mathbb{R}^{e\times e}$ PSD; $\Sigma_{12} \in \mathbb{R}^{e\times f}$ PSD

$\mathbf{a}_2, \mathbf{u}_2 \in \mathbb{R}^d$; $\Sigma_{22} \in \mathbb{R}^{f\times f}$ PSD; $\Sigma_{2,1} \in \mathbb{R}^{f\times e}$ PSD

Predictive density:

$$p(y_{n+1}|x_{n+1},\mathbf{X},\mathbf{y}) = \mathcal{N}(\mu_{n+1},\sigma_{n+1}^2)$$

$$\mu_{n+1} = \mathbf{k}^T\mathbf{C}_n^{-1}\mathbf{y} \quad \sigma_{n+1}^2 = c - \mathbf{k}^T\mathbf{C}_n^{-1}\mathbf{k}$$

## Numerical Estimation Techniques

### Cross-Validation

| | |
|---|---|
| fitted models | $\hat{f}^{-\nu} \in \arg\min_{f\in\mathcal{F}} \frac{1}{\|\mathcal{Z}\setminus\mathcal{Z}_\nu\|}\sum_{i\notin\mathcal{Z}_\nu}(y_i - f(x_i))^2$ |
| pred. error | $\hat{R}^{cv} = \frac{1}{n}\sum_{i\le n}(y_i - \hat{f}^{-\kappa(i)}(x_i))^2$ |
| unbiasedness | $\frac{N(k-1)}{k} \ge m$ (exam2018 k-fold CV) |

### Bootstrap

| | |
|---|---|
| It works if | $R_n^{str}(\hat{F},\hat{F}^*) - R_n^{str}(F,\hat{F}) \overset{P}{\to} 0$ |
| bootst. avg risk: | $\hat{R}^* = \frac{1}{B}\frac{1}{n}\sum_{b=1}^B\sum_{i=1}^n Q(y_i, \hat{f}^{*b}(x_i))$ |
| Sols for overlap: | $\mathcal{C}^{-i} := \{j \in [B] : x_i \notin \mathcal{Z}^{*j}\}$ |
| | $\hat{R}^{(1)} = \frac{1}{n}\sum_{i=1}^n\frac{1}{\|\mathcal{C}^{-i}\|}\sum_{b\in\mathcal{C}^{-i}} l(y_i, f^{*b}(x_i))$ |
| ($\hat{R}^*$ fit on trainset) | $\hat{R}^{.632} = 0.368\hat{R}^* + 0.632\hat{R}^{(1)}$ |
| | $\hat{R}^{(0.632+)} = (1-\hat{w}) * \hat{R}^* + \hat{w}\hat{R}^{(1)}$ |
| $\hat{w} = \frac{0.632}{1-0.368\hat{G}}, \hat{G} =$ | $\frac{\hat{R}^{(1)} - \hat{R}^*}{\hat{\gamma} - \hat{R}^*}, \hat{\gamma} = \frac{1}{n^2}\sum_{i=1}^N\sum_{j=1}^N l(y_i, \hat{f}(x_j))$ |

### Jackknife

$$\hat{S}_{n-1}^{-i}(x_1,...,x_{i-1},x_{i+1},...,x_n) = \hat{S}_{n-1}(x_1,...,x_{i-1},x_{i+1},...,x_n)$$

$$\tilde{S}_n := \frac{1}{n}\sum_{i=1}^n \hat{S}_{n-i}^{-i} \quad bias^{JK} := (n-1)(\tilde{S}_n - \hat{S}_n)$$

(Jackknife) Debiasd estm. $\hat{S}^{JK} = \hat{S}_n - bias^{JK}$

## Tests and criteria

Let $X_1,...,X_n \sim Q(\mathbf{x})$ i.i.d. and $\mathcal{H}_0 : \mathbf{Q} = P_0$ $\mathcal{H}_1 : \mathbf{Q} = P_1$. Test

$$g(x_1,...,x_n) = \begin{cases} 0(accepted) & \frac{P_0(x_1,...,x_n)}{P_1(x_1,...,x_n)} > T \\ 1(rejected) & \frac{P_0(x_1,...,x_n)}{P_1(x_1,...,x_n)} \le T \end{cases}$$

Then $\alpha^* = \mathbb{E}_0[g(x_1,...,x_n)]$ and $\beta^* = 1 - \mathbb{E}_1[g(x_1,...,x_n)]$ Assume that we know the log likelihood function (loss) of the model

| | |
|---|---|
| Bayes Factor | $p(\mathbf{X}\|\mathcal{M}_k)/p(\mathbf{X}\|\mathcal{M}_l)$ $(i)$ |
| $(i) > 1$ take $\mathcal{M}_k$ | $p(\mathbf{X}\|\mathcal{M}_k) = \int p(\mathbf{X}\|\theta_k,\mathcal{M}_k)p(\theta_k\|\mathcal{M}_k)d\theta_k$ |
| BIC (minimise) | $-2\log(\hat{p}(\mathbf{X}\|\hat{\theta}_k,\mathcal{M}_k)) + k'\log n$ |
| Laplace approx. | ($k' = \#$free params in $\mathcal{M}_k$) |
| $\log p(\mathbf{X}\|\mathcal{M}_k) = \log$ | $p(\mathbf{X}\|\hat{\theta}_k,\mathcal{M}_k) - \log(n)k'/2 + O(1)$ |
| MDL | $-\log p(\mathbf{X}\|\theta_k) - \log p(\theta_k)$ |
| AIC | $-2\log(\hat{p}(\mathbf{X}\|\hat{\theta}_k)) + 2k$ |
| KL | $D(p\|\|\hat{p}) = -\int p(x)\log\left(\frac{\hat{p}(x\|\hat{\theta}_k)}{p(x)}\right)dx$ |
| TIC | $-2\log(\hat{p}(\mathbf{X}\|\hat{\theta}_k)) + 2trace[I_1(\theta_k)J_1^{-1}(\theta_k)]$ |

AIC is asymptotically equivalent to LOOCV for ordinary linear regression models.

## Linear Discriminant Functions

| | |
|---|---|
| Gradient Descent | $a_{k+1} = a_k - \eta_k\nabla J(a_k)$ |
| $J(a_{k+1}) \approx J(a_k) + \nabla J^T(a_{k+1} - a_k) + \frac{1}{2}(a_{k+1} - a_k)^T H(a_{k+1} - a_k)$ | |
| $\eta^{OPT} =$ | $\frac{\|\|\nabla J\|\|^2}{\nabla J^T H \nabla J}$ |
| Newton's Rule | $a_{k+1} = a_k - H^{-1}\nabla J(a_k)$ |
| Percep loss | $J(a) = \sum_{\tilde{x}\in\tilde{\mathcal{X}}^{mc}}(-a^T\tilde{x})$ |
| Percep update | $a_{k+1} = a_k + \eta_k\sum_{\tilde{x}\in\tilde{\mathcal{X}}^{mc}}\tilde{x}$ |
| $\gamma = \min_{i\in\tilde{\mathcal{X}}^{mc}}(\hat{a}^T\tilde{x}_i)$ | $\beta^2 = \max_{i\in\tilde{\mathcal{X}}^{mc}}\|\|\tilde{x}_i\|\|^2$ |
| Max steps | $(\gamma)^{-2}\beta^2\|\|\hat{a}\|\|^2$ |

Bayesian view:

`Prior` $P(Y=y)=\pi_y$

`Posterior density` $p(y|x)=\frac{\pi_y p(x|y)}{\sum_z \pi_z p_z(x)}$

$$c(x)=\begin{cases} y & \sum_z L(z,y)p(z|x)=min_{\rho\leq k}\sum L(z,\rho)p(z|x)\leq d \\ \mathcal{D} & else \end{cases}$$

Outlier classi.: $\pi_{\mathcal{O}}p_{\mathcal{O}}(x)\geq \max\{(1-d)p(x), \max_z \pi_z p_z(x)\}$

Fisher's Linear Discriminant Analysis (LDA)

| | |
|---|---|
| `sample avg` | $m_\alpha=\frac{1}{n_\alpha}\sum_{x\in\mathcal{X}_\alpha} x, n_\alpha=\lvert\mathcal{X}_\alpha\rvert$ |
| `projected avg` | $\tilde{m}_\alpha=\frac{1}{n_\alpha}\sum_{x\in\mathcal{X}_\alpha}\mathbf{w}^T x=\mathbf{w}^T m_\alpha$ |
| `class scatter` | $\mathbf{\Sigma}_\alpha=\sum_{x_\alpha\in\mathcal{X}_\alpha}(x-m_\alpha)(x-m_\alpha)^T$ |
| `within scatter` | $\mathbf{\Sigma}_W=\sum_{1\leq\alpha\leq k}\mathbf{\Sigma}_\alpha$ |
| `projected scatter` | $\tilde{\mathbf{\Sigma}}_\alpha=\mathbf{w}^T\mathbf{\Sigma}_\alpha\mathbf{w}$ |
| `Fisher's Separation` | $J(\mathbf{w})=\frac{\mathbf{w}^T(m_1-m_2)(m_1-m_2)^T\mathbf{w})}{\mathbf{w}\mathbf{\Sigma}_W\mathbf{w}}$ |
| yields | $\mathbf{w}\propto\mathbf{\Sigma}_W^{-1}(m_1-m_2)$ |
| `Mean scatter` | $\mathbf{\Sigma}_B=(m_1-m_2)(m_1-m_2)^T$ |
| `result` $\Sigma_W^{-1}\Sigma_B w$ | $=\frac{w^T\Sigma_B w}{w^T\Sigma_W w}w$ |

# Lagrangian Optimization

$$\begin{aligned} &\min f(\mathbf{w}) \quad \mathbf{w}\in\Omega\subseteq\mathbb{R}^d \\ &s.t.\ g_i(\mathbf{w})\leq 0 \quad 1\leq i\leq k \\ &\quad h_j(\mathbf{w})=0 \quad 1\leq j\leq m \end{aligned}$$

$$\mathcal{L}(\mathbf{w},\boldsymbol{\alpha},\boldsymbol{\beta})=f(\mathbf{w})+\sum_{i=1}^k\alpha_i g_i(\mathbf{w})+\sum_{j=1}^m\beta_j h_j(\mathbf{w})$$

$$\frac{\partial\mathcal{L}}{\partial\mathbf{w}}|_{\mathbf{w}=\mathbf{w}^*}=0$$

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}}\theta(\boldsymbol{\alpha},\boldsymbol{\beta}) \text{ with} \theta(\boldsymbol{\alpha},\boldsymbol{\beta})=\inf_{\mathbf{w}}\mathcal{L}(\mathbf{w},\boldsymbol{\alpha},\boldsymbol{\beta})$$

$$s.t.\ \alpha_i\geq 0$$

Duality gap $\Delta:=\mathcal{L}(\mathbf{w}^*,\alpha^*,\beta^*)-\theta(\alpha^*,\beta^*)$

Strong duality, i.e. convex obj. fct $f$ & convex domain, then the duality gap is zero.

KKT Conditions: $f\in C^1$ and $g_i, h_i$ are affine, then $\mathbf{w}^*$ is an optimum if $\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*$ satisfy

$$\frac{\partial\mathcal{L}(\mathbf{w}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)}{\partial\mathbf{w}}=0 \frac{\partial\mathcal{L}(\mathbf{w}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)}{\partial\boldsymbol{\beta}}=0$$

$$\alpha_i^* g_i(\mathbf{w}^*)=0, g_i(\mathbf{w}^*)\leq 0, \alpha_i^*\geq 0$$

# SVM

Soft Margin Geometric problem formulation

| | |
|---|---|
| `Primal` | $\min_{\mathbf{w},\boldsymbol{\xi}}\frac{1}{2}\mathbf{w}^T\mathbf{w}+C\sum_{i=1}^n\xi_i$ |
| | $z_i(\mathbf{w}^T\mathbf{y}_i+w_0)\geq 1-\xi_i \quad \xi_i\geq 0$ |
| `Dual` | $\max_{\boldsymbol{\alpha}}\sum_{i\leq n}\alpha_i-\frac{1}{2}\sum_{i\leq n}\sum_{j\leq n}\alpha_i\alpha_j z_i z_j\mathbf{y}_i^T\mathbf{y}_j^T$ |
| | $C\geq\alpha_i\geq 0, \sum_{i\leq n}z_i\alpha_i=0$ |
| `Solution` | $w_0^*=(\max_{i:z_i=-1}\mathbf{w}^{*T}\mathbf{y}_i+\min_{i:z_i=1}\mathbf{w}^{*T}\mathbf{y}_i)/2$ |
| | $\mathbf{w}^*=\sum_{i\in SV}\alpha_i^* z_i\mathbf{y}_i$ |
| | $g^*(\mathbf{y})=\sum_{i\in SV}z_i\alpha_i^*\mathbf{y}_i^T\mathbf{y}+w_0^*$ |

By the KKT condition, $\xi_i(\alpha_i-C)=0$, non-zero slack variable can only occur if $\alpha_i=C$.

The optimal margin is given by: $\mathbf{w}^\top\mathbf{w}=\sum_{i\in SV}\alpha_i^*$

Multi-class SVM: $\mathbf{w}^T=(\mathbf{w}_1^T,...,\mathbf{w}_n^T)$.

| | |
|---|---|
| `Primal` | $\min_{\mathbf{w},\boldsymbol{\xi}}\frac{1}{2}\mathbf{w}^T\mathbf{w}+C\sum_{i\leq n}\xi_i$ |
| $\xi_i\geq 0$ | $(\mathbf{w}_{z_i}^T\mathbf{y}_i+w_{z_i,0})-\max_{z\neq z_i}(\mathbf{w}_z^T\mathbf{y}_i+w_{z,0})\geq 1-\xi_i$ |

Structured SVM:

| | |
|---|---|
| `Primal` | $\min_{\mathbf{w},\boldsymbol{\xi}}\frac{1}{2}\mathbf{w}^T\mathbf{w}+C\sum_{i\leq n}\xi_i$ s.t. $\xi_i\geq 0$ |
| $\mathbf{w}^T\mathbf{\Psi}(z_i,\mathbf{y}_i)$ | $-\max_{z\neq z_i}[\Delta(z,z_i)+\mathbf{w}^T\mathbf{\Psi}(z,\mathbf{y}_i)]\geq-\xi_i$ |
| $(\mathbf{w}^T\mathbf{\Psi}(z_i,\mathbf{y}_i)$ | $-\mathbf{w}^T\mathbf{\Psi}(z,\mathbf{y}_i)\geq\Delta(z,z_i)-\xi_i \quad \forall i,\forall z\neq z_i)$ |
| `Dual` $\min_{\mathbf{w},\boldsymbol{\xi}}$ | $-\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n\sum_{z_k\in\mathbb{K}}\alpha_{ik}\alpha_{jk}\Psi_i(z_k)^\top\Psi_j(z_k)$ |
| | $+\sum_{i=1}^n\sum_{z_k\in\mathbb{K}}\alpha_{ik}\Delta_i(z_k)$ |
| s.t. | $C\geq\sum_{z_k\in\mathcal{K}}\alpha_{ik}\geq 0,\ \alpha_{ik}\geq 0, \forall i,\forall k$ |
| `Prediciton` | $h(y)=\arg\max_{z\in\mathbb{K}}\left\{\mathbf{w}^\top\psi(z,\mathbf{y})\right\}$ |

# Ensemble

If we combine different regressors: $V[\hat{f}(x)]\approx\frac{\sigma^2}{B}$

$bias[\hat{f}(x)]=\frac{1}{B}\sum_{i=1}^B bias[\hat{f}_i(x)]$

**Boosting**: Weighted models and weighted training data instead of bootstrapping.

$$\epsilon_b\leftarrow\sum_{i=1}^n w_i^{(b)}\mathbb{I}_{\{c_b(x_i)\neq y_i\}}/\sum_{i=1}^n w_i^{(b)}$$

$$\alpha_b\leftarrow\log\frac{1-\epsilon_b}{\epsilon_b}=\log\left(\frac{p(y=1|x)}{p(y=-1|x)}\right) \text{ (log-odds ratio)}$$

$$\forall i\ w_i\leftarrow w_i\exp(\alpha_b\mathbb{I}_{\{c_b(x_i)\neq y_i\}})$$

$$\hat{c}_B(x)=sign\left(\sum_{b=1}^B\alpha_b c_b(x)\right)$$

`avg exp loss` $=\frac{1}{N}\sum_{i=1}^N\exp(-y_i\hat{c}_B(x_i))$

$Err_{\text{AdaBoost}}=\exp(-h(x)sign(\sum_{b=1}^B\alpha_b y_b(x)))$

# PAC Learning

| | |
|---|---|
| `error` | $error(h)=P_{x\sim\mathcal{D}}[c(x)\neq h(x)]$ |
| $(\epsilon,\delta)$ criterion: | $P_{X,Y}[R(\hat{c}_n)\leq R(c^{Bayes})+\epsilon]>1-\delta$ |
| **Strong PAC L.**: | holds for arbitrarily small $\epsilon$ |
| **Weak PAC L.**: | non-trivially large $\epsilon$ |
| `PAC learnability` | $P[\mathcal{R}(\hat{c}_n)\leq\varepsilon]\geq 1-\delta$ |
| `efficiently` | $\mathcal{A}$ runs in poly time in $\frac{1}{\varepsilon}$ and $\frac{1}{\delta}$ |

Results:

| | |
|---|---|
| $R(\hat{c}_n^*)-\inf_{c\in\mathcal{C}}R(c)$ | $\leq 2\sup_{c\in\mathcal{C}}\lvert\hat{R}_n(c)-R(c)\rvert$ |
| $P[\sup_{c\in\mathcal{C}}\lvert\hat{R}_n(c)-R(c)\rvert>\epsilon]$ | $\leq 2N\exp(-2n\epsilon^2)$ |
| Implying: $R(c)\leq\hat{R}_n(c)$ | $+\sqrt{(\log N-\log(\delta/2))/2n}$ |
| $X$ shattered by $\mathcal{A}$ if | |
| $\{X\cap A\lvert A\in\mathcal{A}\}$ | contains all subsets of $X$ |
| `VC Dim` of $\mathcal{A}=\max\{n:\exists X$ | $s.t.\lvert X$ shattered by $\mathcal{A},\lvert X\rvert=n\}$ |
| `score` | $score(\mathcal{A},X)=\lvert\{X\cap\mathcal{A}\lvert A\in\mathcal{A}\}\rvert$ |
| `shattering coeff` | $s(\mathcal{A},n)=\max_{X:\lvert X\rvert=n}score(\mathcal{A},X)$ |
| If $V_{\mathcal{A}}>2$ ($V_{\mathcal{A}}=$ VC dim. of | $\mathcal{A}$): $s(\mathcal{A},n)\leq n^{V_{\mathcal{A}}}$ |
| $P[R(c_n^*)-inf_{c\in\mathcal{C}}R(c)>\epsilon]$ | $\leq 8s(\mathcal{A},n)\exp(-n\epsilon^2/32)$ |

# Non Paramteric Bayesian Methods

| | |
|---|---|
| `Beta function` | $B(a,b)=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, a,b>0$ |
| | $\Gamma(a)=\int_0^\infty e^{-x}x^{a-1}dx$ |
| $Beta(x\lvert a,b)=$ | $\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, x\in[0,1]$ |
| $Dir(\mathbf{x}\lvert\boldsymbol{\alpha})=$ | $\frac{\prod_{k=1}^n x_k^{\alpha_k-1}}{B(\alpha)}$ |
| `Finite Gaussian Mix` | $p(\mathbf{x}_i\lvert\boldsymbol{\theta})=\sum_{k=1}^K\rho_k\mathcal{N}(\mathbf{x}_i\lvert\mu_k,\sigma_k)$ |

Stick breaking process (GEM distribution):

$\beta_k\sim Beta(1,\alpha) \quad \rho_k=\beta_2(1-\sum_{i=1}^{k-1}\rho_i), \quad k=1,2,...$

**Chinease Restaurant Process**

$P(\mathcal{P})=\frac{\alpha^{\lvert\mathcal{P}\rvert}}{\alpha^{(n)}}\prod_{\tau\in\mathcal{P}}(\lvert\tau\rvert-1)! \quad \mathbb{E}[\#k]=\sum_{i=1}^N\frac{\alpha}{\alpha+i}\sim O(\alpha\log N)$

$$\text{P[Customer } n+1 \text{ joins table } \tau\in\mathcal{P}\cup\{\emptyset\}|\mathcal{P}]=\begin{cases}\frac{\lvert\tau\rvert}{\alpha+n} & \tau\in\mathcal{P}\\ \frac{\alpha}{\alpha+n} & ow.\end{cases}$$

Dirichlet Mixture Model

| | |
|---|---|
| `Base Measures` | $\mu_k\sim\mathcal{N}(\mu_0,\sigma_0)$ |
| `cluster prob.` | $\rho=(\rho_1,\rho_2,...)\sim GEM(\alpha)$ |
| `Category assignment` | $z_i\sim Categorical(\rho)$ |
| `Data Sample` | $x_i\sim\mathcal{N}(\mu_{z_i},\sigma)$ |

De Finetti's Theorem

$p(X_1,...,X_n)=\int\prod p(x_i|G)dP(G)$

Gibbs Sampling

$$p(z_i=k|\mathbf{z}_{-i},\mathbf{x},\alpha,\boldsymbol{\mu})\propto\underbrace{p(z_i=k|\mathbf{z}_{-i},\alpha)}_{Prior}\underbrace{p(x_i|\mathbf{x}_{-i},z_i=k,\mathbf{z}_{-i},\boldsymbol{\mu})}_{Likelihood}$$

$$p(z_i=k|\mathbf{z}_{-i},\mathbf{x},\alpha,\boldsymbol{\mu})=\begin{cases}\frac{N_{k,-i}}{\alpha+N-1}p(x_i|\mathbf{x}_{-i,k},\boldsymbol{\mu}) & \text{For existing } k\\ \frac{\alpha}{\alpha+N-1}p(x_i|\boldsymbol{\mu}) & \text{Otherwise}\end{cases}$$

$$p(x_i,\mathbf{x}_{-i,k}|\boldsymbol{\mu})=\int p(x_i|\boldsymbol{\mu_k})\left(\prod_{j\neq i}p(x_j|\boldsymbol{\mu_k})\right)p(\boldsymbol{\mu}_k|\mu_0,\sigma_0)d\boldsymbol{\mu}_k$$

# Gaussian-mixtures and EM estimation

| | |
|---|---|
| `Parameters` | $\theta=\{\pi_c,\mu_c,\sigma_c^2\}_{c=1}^k$ |
| `k Gaussian Mix.` | $P(x_i\lvert\theta)=\sum_{c=1}^k\pi_c P(x_i\lvert c,\mu_c,\sigma_c^2)$ |
| | $\sum_{c=1}^k\pi_c=1$ |
| `log likelihood` | $L(\mathcal{X}\lvert\theta)=\log P(\mathcal{X}\lvert\theta)=\sum_{i=1}^n\log P(x_i\lvert\theta)$ |
| | $=\sum_{i=1}^n\log\sum_{c=1}^k\pi_c P(x_i\lvert c,\mu_c,\sigma_c^2)$ |

Define binary latent variables $M_{ic}\in\{0,1\}$ where $M_{ic}$ indicates that $x_i$ is generated by component $c$. `log likelihood`

$L(\mathcal{X},M|\theta)=\log\prod_{i=1}^n\prod_{c=1}^k(\pi_c P(x_i|c,\mu_c,\sigma_c^2))^{M_{ic}}$

$L(\mathcal{X},M|\theta)=\sum_{i=1}^n\sum_{c=1}^k M_{ic}\log(\pi_c P(x_i|c,\mu_c,\sigma_c^2))$

`Expectation over the latent va.` $\left(\gamma_{ic}:=E_{M|\mathcal{X},\theta}[M_{ic}]\right)$

$Q(\theta):=E_{M|\mathcal{X},\theta}[L(\mathcal{X},M|\theta)]=\sum_{i=1}^n\sum_{c=1}^k\gamma_{ic}\log(\pi_c P(x_i|c,\mu_c,\sigma_c^2))$

`EM-estimation algo`
- **E-step** compute $\gamma_{ic}$, $\theta$ const
- **M-step** compute $\theta$, $\gamma_{ic}$ const

**E-step:** $E_{M|\mathcal{X},\theta}[M_{ic}]=1\cdot P(M_{ic}=1|x_i,\theta)+0\cdot P(M_{ic}=0|x_i,\theta)$

$=P(M_{ic}=1|x_i,\theta)=P(c|x_i,\theta)=\frac{P(x_i|c,\theta)P(c|\theta)}{P(x_i|\theta)}$

$=\frac{\pi_c P(x_i|c,\mu_c,\sigma_j^2)}{\sum_{j=1}^k\pi_j P(x_i,|j,\mu_j,\sigma_j^2)}$

**M-step:**

(i) $\mu_c$ `from` $\arg\max_\theta Q(\theta)$:

$\frac{\partial}{\partial\mu_c}Q(\theta)=0 \quad\Longrightarrow\quad \mu_c=\frac{\sum_{i=1}^n\gamma_{ic}x_i}{\sum_{i=1}^n\gamma_{ic}}$

(ii) $\sigma_c$ `from` $\arg\max_\theta Q(\theta)$:

$\frac{\partial}{\partial\sigma_c}Q(\theta)=0\Longrightarrow \quad \sigma_c^2=\frac{\sum_{i=1}^n\gamma_{ic}(x_i-\mu_c)^2}{\sum_{i=1}^n\gamma_{ic}}$

(ii) $\pi_c$ `from` $\arg\max_\theta Q(\theta)$: constraint $\sum_c\pi_c=1$

$\mathcal{L}(\theta,\lambda)=-Q(\theta)+\lambda(\sum_{c=1}^k\pi_c-1) \Rightarrow\frac{\partial}{\partial\pi_c}\mathcal{L}(\theta,\lambda)=0$

$\Leftrightarrow\sum_{i=1}^n\gamma_{ic}=\lambda\pi_c\Leftrightarrow\sum_{c=1}^k\sum_{i=1}^n \quad \gamma_{ic}=\lambda\sum_{i=1}^c\pi_c$

$\Leftrightarrow\sum_{i=1}^n\sum_{c=1}^k\gamma_{ic}=\sum_{i=1}^n 1=\lambda \quad\Leftrightarrow\quad \pi_c=\frac{\sum_{i=1}^n\gamma_{ic}}{n}$