Disclaimer

This document is an exam summary and follows the given material of the lecture *Advanced Machine Learning*. Its contribution is a short summary that contains the most important concepts, formulas and algorithms. Due to curriculum content updates, some content may not be relevant to future versions of the course.

I do not guarantee the accuracy or completeness, nor is this document endorsed by the instructors. Any errors that are pointed out to me are welcome. The complete LaTeX source code can be found at `https://github.com/tstreule/eth-cheat-sheets`.

Advanced Machine Learning

Timo Streule, tstreule@ethz.ch - *18.01.2022*

1 Basics

- General p-norm: $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$
- Taylor: $f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \dots$
  - $f(\boldsymbol{a}) + \frac{\partial f(x)}{\partial x}\Big|_a - \frac{1}{2}(x-a)^\top \left(\frac{\partial^2 f(x)}{\partial x \partial x^\top}\right)\Big|_a (x-a)$
  - Power series of exp.: $\exp(x) := \sum_{k=0}^\infty \frac{x^k}{k!}$
- Entropy: $H(X) = \mathbb{E}_X[-\log \mathbb{P}(X=x)]$
- Diverg.: $D_{KL}(P \,\|\, Q) = \sum_{x\in\mathcal{X}} P(x)\log\left(\frac{P(x)}{Q(x)}\right) \ge 0$
- $1-z \le \exp(-z)$
- Cauchy-Schwarz: $|\mathbb{E}[X,Y]|^2 \le \mathbb{E}[X^2]\mathbb{E}[Y^2]$
- Jensen, $f(X)$ convex: $f(\mathbb{E}[X]) \le \mathbb{E}[f(X)]$

1.1 Probability / Statistics

- Gaussian: $\mathcal{N}(x\,|\,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac12\frac{(x-\mu)^2}{\sigma^2}\right)$

  $\mathcal{N}(x|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp\left(-\frac12(x-\mu)^\top\Sigma^{-1}(x-\mu)\right)$

  $X\sim\mathcal{N}(\mu,\Sigma),\ Y=A+BX \Rightarrow Y\sim\mathcal{N}(A+B\mu, B\Sigma B^\top)$
- Binom.: $f(k,n;p) = \mathbb{P}(X=k) = \binom{n}{k}p^k(1-p)^{n-k}$
- $\mathbb{V}[X] = \mathbb{E}[(X-\mathbb{E}[X])^2] = \mathbb{E}[X^2]-\mathbb{E}[X]^2$

  $\mathbb{V}[X+Y] = \mathbb{V}[X]+\mathbb{V}[Y]+2\mathbb{C}ov(X,Y)$
- $\mathbb{C}ov(X,Y) = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]$

  $= \mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]$

  $\mathbb{C}ov(aX,bY) = ab\,\mathbb{C}ov(X,Y)$

1.2 Calculus

- $\int uv'\,dx = uv - \int u'v\,dx$ • $\frac{\partial}{\partial x}\frac{g}{h} = \frac{g'h}{h^2} - \frac{gh'}{h^2}$
- $\frac{\partial}{\partial x}(b^\top A x) = A^\top b$ • $\frac{\partial}{\partial x}(b^\top x) = \frac{\partial}{\partial x}(x^\top b) = b$
- $\frac{\partial}{\partial X}(c^\top X^\top b) = bc^\top$ • $\frac{\partial}{\partial X}(c^\top X b) = cb^\top$
- $\frac{\partial}{\partial x}(x^\top A x) = (A^\top + A)x \overset{A\text{ sym.}}{=} 2Ax$
- $\frac{\partial}{\partial X}Tr(X^\top A) = A$ • Tr. trick: $x^\top A x \overset{\text{inner prod.}}{=} Tr(x^\top A x) \overset{\text{cyclic permut.}}{=} Tr(xx^\top A) = Tr(Axx^\top)$
- $|X^{-1}| = |X|^{-1}$ • $\frac{\partial}{\partial X}\log|X| = X^{-\top}$ • $\frac{d}{dx}|x| = \frac{x}{|x|}$
- $\frac{\partial}{\partial x}\|x\|_2 = \frac{\partial}{\partial x}(x^\top x) = 2x$ • $\frac{\partial}{\partial x}\|x-b\|_2 = \frac{x-b}{\|x-b\|_2}$
- $\frac{\partial}{\partial x}\|x\|_1 = sgn(x)$ $\quad sgn(x)\in\{\pm1\}^p$ is row-wise
- $\sigma(x) = \frac{1}{1+\exp(-x)} \Longrightarrow \nabla\sigma(x) = \sigma(x)(1-\sigma(x))$
- $\tanh x = \frac{2\sinh x}{2\cosh x} = \frac{e^x-e^{-x}}{e^x+e^{-x}}$ • $\nabla\tanh x = 1-\tanh^2 x$

2 Density Estimation

Bayesianism: Define prior $P(\theta)$, define likelihood $P(X\,|\,\theta)$, compute posterior $P(\theta\,|\,x_{1\dots n})$.

Bayes: $P(\theta\,|\,X) = \frac{P(X|\theta)P(\theta)}{P(X)}$, $P(X) = \sum_\theta P(X|\theta_i)P(\theta_i)$

Frequentism: Define param. model $P(Y\,|\,X,\theta)$, compute likelihood of data $P(X,Y\,|\,\theta)$ and compute $\hat\theta_{MLE}$ via $\arg\max_\theta$ of likelihood.

2.1 Estimation - MLE Properties

Consistency: $\forall\epsilon>0,\ \mathbb{P}\{|\hat\theta_n - \theta^*|>\epsilon\} \overset{n\to\infty}{\longrightarrow} 0$

Equivariance: If $\hat\theta_n$ is MLE of θ, then $g(\hat\theta_n)$ is MLE of $g(\theta)$.

Asympt. normality:
$\sqrt{n}(\hat\theta_n-\theta^*) \to \mathcal{N}(0, J^{-1}(\theta^*)I_n(\theta^*)J^{-1}(\theta^*))$

Asympt. efficiency: $\hat\theta_n$ minimises $\mathbb{E}[(\hat\theta_n-\theta^*)^2]$ as $n\to\infty$, i.e. $\mathbb{E}[(\hat\theta_n-\theta^*)^2] \overset{n\to\infty}{=} \frac{1}{I_n(\theta^*)}$ (Rao Cr.)

Among all consistent estimators $\hat\theta_n$ has *smallest variance*: $\lim_{n\to\infty}(\mathbb{V}[\hat\theta_n]I_n(\theta^*))^{-1} = 1$

2.2 Rao Cramer inequality — all $\mathbb{E}$ w.r.t. $P(x\,|\,\theta^*)$

Score func.: $\Lambda = \frac{\partial\log\mathbb{P}(x|\theta)}{\partial\theta}$, $\mathbb{E}[\Lambda]=0$

Fisher info.: $I_n(\theta) = \mathbb{V}[\Lambda]$

$J(\theta) = \mathbb{E}[\Lambda^2] = -\mathbb{E}\left[\frac{\partial^2\log\mathbb{P}(x|\theta)}{\partial\theta\partial\theta^\top}\right] = -\mathbb{E}\left[\frac{\partial\Lambda}{\partial\theta}\right]$

General bound: $\mathbb{E}[(\hat\theta_n-\theta^*)^2] \ge \frac{(1+\frac{\partial}{\partial\theta}b_{\hat\theta})^2}{\mathbb{E}[\Lambda^2]} + b_{\hat\theta}^2$

Unbiased case: $\mathbb{E}[(\hat\theta_n-\theta^*)^2] = \mathbb{V}[\hat\theta_n] \ge \frac{1}{I_n(\theta^*)}$

Tradeoff: $\mathbb{E}[(\hat\theta_n-\theta^*)^2] = \mathbb{V}[\hat\theta_n] + bias^2(\hat\theta_n)$

Bias: $bias(\hat\theta_n) \equiv b_{\hat\theta}(\theta^*) = \mathbb{E}[\hat\theta_n]-\theta^* \overset{\text{unbiased}}{=} 0$

3 (Linear) Regression — model: $\hat y = X\beta$

Assuming $X^\top X$ non-singular.

Bayesian view: $(Y\,|\,X,\beta)\sim\mathcal{N}(x^\top\beta, \sigma^2\mathbb{I})$

Distrib. of estimator $\hat\beta_{LS}\sim\mathcal{N}(\beta,(X^\top X)^{-1}\sigma^2)$

Ridge: $\epsilon_{RSS}(\beta,\lambda) = (y-X^\top\beta)^\top(y-X^\top\beta)+\lambda\beta^\top\beta$

$\hat\beta = (X^\top X+\lambda\mathbb{I})^{-1}X^\top y$, prior: $\beta\sim\mathcal{N}(0,\frac{\sigma^2}{\lambda}\mathbb{I})$

(Ridge) Shrinkage: Decompose $X = U\,D\,V^\top$

$X\hat\beta = U\,D(D^2+\lambda\mathbb{I})^{-1}DU^\top y = \sum_{j\le d} u_j\frac{d_j^2}{d_j^2+\lambda}u_j^\top y$

Lasso: $\hat\beta = \arg\min_\beta\sum_{i\le n}(y_i-x_i^\top\beta)^2+\lambda\|\beta\|_1$

(no closed form), prior: $p(\beta_i) = \frac{\lambda}{4\sigma^2}\exp(-|\beta_i|\frac{\lambda}{2\sigma^2})$

Bias-variance: $\mathbb{E}_D[\mathbb{E}_{Y|X=x}[(\hat f(x)-Y)^2]]$

$= \mathbb{E}_D[(\hat f(x)-\mathbb{E}_D[\hat f(x)])^2] + (\mathbb{E}_D[\hat f(x)]-\mathbb{E}[Y|X=x])^2$

$+\mathbb{E}[(Y-\mathbb{E}[Y|X=x])^2] = variance + bias^2 + noise$

Gauss-Markov Theorem: For any linear estimator $\tilde\theta = c^\top y = a^\top(\hat\beta+Dy)$ that is unbiased for $a^\top\beta$, it holds: $\mathbb{V}[a^\top\hat\beta] \le \mathbb{V}[c^\top y]$.

Among all linear **u**-estimators, $\hat\beta_{LS}$ minimises the gen. error! What about **biased** estimators? We ↗ bias a bit in the hope that the variance ↘.

Combining Regressors: $\hat f(x) := \frac1B\sum_{i\le B}\hat f_i(x)$

$bias[\hat f(x)] = \frac1B\sum bias[\hat f_i(x)]$

$\mathbb{V}[\hat f] = \frac{1}{B^2}\sum\mathbb{V}_D[\hat f_i] + \frac{1}{B^2}\sum\sum_{i\ne j}\mathbb{C}ov(\hat f_i,\hat f_j) \approx \frac{\sigma^2}{B}$

4 Gaussian Processes

4.1 Bayesian Linear Regression

Model: $y = X^\top\beta+\epsilon$, with $\epsilon\sim\mathcal{N}(\epsilon\,|\,0,\sigma^2\mathbb{I})$

Likelihood: $P(y\,|\,X,\beta,\sigma) = \mathcal{N}(y\,|\,X^\top\beta,\sigma^2\mathbb{I})$

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

(Ridge regr. if $\Lambda=\lambda\mathbb{I}$ and $\sigma=1$)

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

with $\mu_\beta = (X^\top X+\sigma^2\Lambda)^{-1}X^\top y$

and $\Sigma_\beta = \sigma^2(X^\top X+\sigma^2\Lambda)^{-1}$

$\Longrightarrow y\sim\mathcal{N}(y\,|\,0, X\Lambda^{-1}X^\top+\sigma^2\mathbb{I})$ using $\mathbb{E}_{\beta,\epsilon}[\cdot]$

kernel $k(x_i,x_j) := x_i^\top\Lambda^{-1}x_j$

4.2 Gaussian Process

$y\sim\mathcal{N}(y\,|\,m(X),K(X,X)+\sigma^2\mathbb{I})$

$\begin{bmatrix}y\\y_{n+1}\end{bmatrix}\sim\mathcal{N}\left(\begin{bmatrix}y\\y_{n+1}\end{bmatrix}\middle|\begin{bmatrix}m(X)\\m(x_{n+1})\end{bmatrix},\begin{bmatrix}C_n & k\\k^\top & c\end{bmatrix}\right)$

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

with $\mu_{n+1} = m(x_{n+1})+k^\top C_n^{-1}(y-m(X))$

and $\sigma^2_{n+1} = c-k^\top C_n^{-1}k$

where $K=k(X,X)$, $k=k(x_{n+1},X)$,

$C_n = K+\sigma^2\mathbb{I}$, $c=k(x_{n+1},x_{n+1})+\sigma^2$

4.3 Kernels — scalar product $K_{ij}=k(x_i,x_j)$

Valid kernel: must be **symmetric** and **p.s.d.** ($x^\top Kx\ge0\ \forall x$ *or* pos. eigenvalues *or* pos. principal minors). Must have a (pot. ∞-dim.) **feature vector** ϕ s.t. $k(x,x') = \phi(x)^\top\phi(x')$.

Common kernels:

Linear: $x^\top x'$

Polynomial: $(x^\top x'+1)^p$, $p\in\mathbb{N}$

RBF (Gaussian): $\exp(-\|x-x'\|_2^2/h^2)$

Sigmoid: $\tanh(\kappa\cdot x^\top x'-b)$

Kernel construction: • k_1+k_2 • $c\cdot k_1$, $c>0$
- $k_1\cdot k_2$ • $f(x)k_1(x,x')f(x')$
- $k(\phi(x),\phi(x'))$ with $\phi:\mathcal{X}\to\mathbb{R}^d$
- $g(k_1)$ with g : exp. *or* polyn. w/ <u>all</u> pos. coeff.

5 Linear Classification — $y,z\in\{\pm1\}$, $z\equiv c(x)$

(1) Prob. gener. $p(x,y)$ +outlier det.

(2) Prob. discr. $p(y\,|\,x)$ +deg. of belief

(3) Purely discr. $c: X\to y$ +easiest

Loss Functions: $\mathcal{L}(y,z)$ $\quad z := w^\top x$

$\mathcal{L}^{CE} = -[y'\log z' + (1-y')\log(1-z')]$

$\mathcal{L}^{0/1} = \mathbb{I}[sign(z)\ne y]$

$\mathcal{L}^{hinge} = \max(0,1-yz)$ for SVM's

$\mathcal{L}^{percep} = \max(0,-yz)$ for

$\mathcal{L}^{logistic} = \log(1+\exp(-yz))$

$\mathcal{L}^{exp} = \exp(-yz)$ for AdaBoost

CE (log loss): $y'=(1+y)/2$, $z'=(1+z)/2$

5.1 Linear Discriminant Analysis (1)

Assume $Y\sim Ber(\beta)$, $P(X|Y{=}i) = \mathcal{N}(\mu_i,\Sigma_i)$.

$\Rightarrow P(y_i\,|\,x_i) = \sigma(\cancel{x_i^\top W x_i} + w^\top x_i + w_0)$ $\quad$ if $\Sigma_0=\Sigma_1$

Min. gener. error.: $\min_f\mathbb{E}_{X,Y}[\mathcal{L}(y,c(x))]$

$\rightsquigarrow c^*(x) = \arg\min_c\sum_y p(y\,|\,x)\mathcal{L}(y,c(x))$

5.2 Prob. discr. approach (2)

Assume $P(y{=}1\,|\,x_i,w) = \sigma(w^\top x)$, $\Longrightarrow L(w)$ via $P(X,Y\,|\,w) = \prod_i P(y_i\,|\,x_i,w)\underbrace{P(x_i\,|\,w)}_{const\text{ w.r.t. }w}$

$\propto\prod_i\sigma(w^\top x_i)^{y_i}(1-\sigma(w^\top x_i))^{1-y_i}$

Note: w^* intractable but diff'able → GD!

5.3 Purely discriminative (3)

Perceptron: $f(x) = sgn(w^\top x)$

Loss: $L(w) = \sum_{i:\text{misclass.}}(-y_i w^\top x_i)$ use (S)GD

Converges if data is linearly separable, and $\eta(k)\ge0$, $\sum_k\eta(k)\to\infty$, $\sum_k\eta^2(k)<\infty$.

Gradient Descent: $NL(w) := -L(w)$

$w^{(k+1)} \leftarrow w^{(k)} - \eta(k)\cdot\nabla_w NL(w^{(k)})$

Opt. learning rate: $\eta(k) = \arg\min_\eta NL(w^{(k+1)})$

(Taylor & $\frac{\partial}{\partial\eta(k)}\overset{!}{=}0$) $= \frac{\|\nabla NL(w^{(k)})\|^2}{\nabla NL(w^{(k)})^\top H_{NL}(w^{(k)})\nabla NL(w^{(k)})}$

Newton's Method: $w^{(k+1)}\leftarrow\arg\min_w NL(w)$

(Taylor & $\frac{\partial}{\partial w}\overset{!}{=}0$) $= w^{(k)} - H_{NL}^{-1}(w^{(k)})\nabla NL(w^{(k)})$

Fisher's LDA:

$J(w) = \frac{w^\top\Sigma_B w}{w^\top\Sigma_W w} \overset{(*)}{\longrightarrow} w^*\propto\Sigma_W^{-1}(\bar x_1-\bar x_2)$

$*: \frac{\partial J(w)}{\partial w}\overset{!}{=}0 \rightsquigarrow (\cancel{w^\top\Sigma_B w})\Sigma_W w = (\cancel{w^\top\Sigma_W w})\Sigma_B w$

$\Sigma_B = (\bar x_1-\bar x_2)(\bar x_1-\bar x_2)^\top$ between-class covariance

$\Sigma_W = \sum_k\sum_{x\in\mathcal{C}_k}(x-\bar x_k)(x-\bar x_k)^\top$ within-class covariance

6 Support Vector Machine (SVM)

Primal (soft margin): $\min_{w,w_0,\xi}\frac12\|w\|^2+C\sum_i\xi_i$

s.t. $y_i(w^\top x_i+w_0)\ge1-\xi_i$ and $\xi_i\ge0$

↪ intractable if $\varphi(x_i)$ instead of x_i

↪ $\xi_i=0$ means x_i was <u>not</u> neglected

Dual: $\max_\alpha\sum_i\alpha_i - \frac12\sum_{(i,j)}\alpha_i\alpha_j y_i y_j x_i^\top x_j$

s.t. $0\le\alpha_i\le C$; $\sum_i\alpha_i y_i = 0$

↪ solve α via quadratic optimisation

Optimal hyperplane: $w^* = \sum_i\alpha_i^* y_i x_i$

↪ $\alpha_i^*\ne0$ only for *support vectors*

Optimal slack: $\xi_i^* = \max(0,1-y_i(w^{*\top}x_i+w_0^*))$

6.1 Structural SVMs

$\min_{w,\xi}\frac12\|w\|^2+\frac{C}{n}\sum_{i\le n}\xi_i$ s.t. $\xi_i\ge0$ and $\forall y'\ne y_i$:

$w^\top\Psi(x_i,y_i)\ge\Delta(y_i,y')+\underbrace{w^\top\Psi(x_i,y')}_{\hookrightarrow\text{ mislabelings}}-\xi_i-\epsilon$

Ψ : *joint-feature map*; Δ : loss / class dissimilarity func.; $w^\top\Psi(x,y)$: compatibility score btw. x and y; ϵ : tolerance / universal slack variable.

Prediction: $c(x) = \arg\max_y w^\top\Psi(x,y)$

Note: For optimal w^*, ξ^*, emp. risk$(w^*)\le\frac1n\sum_i\xi_i^*$

Training: Start without any constraints.

In each iteration, add for each (x_i, y_i) the constraint with $y' \neq y_i$ that is the "most violated" and solve again with quadr. optimisation.

7 Ensemble Methods

7.1 Bagging (Bootstrap aggregation)

1. Draw M bootstrap sets $Z'_1, \dots, Z'_M$
2. Train M base models $b^{(1)}, \dots, b^{(M)}$
3. Aggregate: $\hat{b}^{(M)}(\mathbf{x}) = \begin{cases} \frac{1}{M}\sum_{t \leq M} b^{(t)}(\mathbf{x}) & \text{regr.} \\ \text{sign}(\sum_t b^{(t)}(\mathbf{x})) & \text{class.} \end{cases}$

Why it works: Small *variance* (weak learners), small *covariance* (almost indep. since $Z'_i \neq Z'_j$).

For finite range y and large enough M:

$$\mathbb{E}_{\substack{Y|X \\ Z,Z'}}\left[(y - \hat{b}^{(M)}(\mathbf{x}))^2\right] \leq \mathbb{E}_{\substack{Y|X \\ Z,Z'}}\left[(y - b(\mathbf{x}))^2\right]$$

Random Forest: At each splitting step, u.a.r. choose m of p features and split only one (best) feature. $\to$ reduce *correlation* between trees.

Validation: *Out-of-bag error* $\to$ validate each $\mathbf{x}_i$ with trees that didn't use it for training.

7.2 Boosting

Sequentially train weak learners on all data, but $\nearrow$ weight of misclass. samples ($\searrow$bias).

AdaBoost: Stat. learning (*forward stagewise additive modeling*) with **exp. loss**, trains max-margin ($= y_i \hat{b}(\mathbf{x}_i)$), self-avg. and interpolating ($\searrow$overfitting) classifiers.

[Init]: $\hat{b}^{(0)} \leftarrow 0$, $w_i \leftarrow 1/n \; \forall i \leq n$
for $t = 1 \dots M$:
[Train]: $b^{(t)} = \arg\min_b \mathcal{L}^w(b) = \sum_i w_i \mathbb{I}\{b(\mathbf{x}_i) \neq y_i\}$
[Eval]: $\text{err}_t = \mathcal{L}^w(b^{(t)})$
[Aggr]: $\hat{b}^{(t)} = \hat{b}^{(t-1)} + \alpha_t b^{(t)}$; $\alpha_t = \frac{1}{2}\log(\frac{1}{\text{err}_t} - 1)$
[Reweight]: $w_i = w_i \cdot \exp(\alpha_t \mathbb{I}\{b^{(t)}(\mathbf{x}_i) \neq y_i\})$ normalize!
Return $\hat{b}^{(M)}(\mathbf{x}) = \text{sign}(\sum_t \alpha_t b^{(t)}(\mathbf{x}))$

8 Deep Learning

Sigmoid: $\sigma(x) = \frac{1}{1+\exp(-x)} = \frac{e^x}{e^x+1}$
$\sigma'(x) = \sigma(x)(1-\sigma(x)) = \sigma(x)\sigma(-x)$

Softmax: $y_i \propto \exp(\beta z_i)$

Backpropagation: Gradient: $\frac{\partial \ell}{\partial w_{jk}} = \delta_j^{(l)} v_k^{(l-1)}$

Error signal for unit k on layer l:
$\delta^{(L)} = [\cdots \delta_k^{(L)} \cdots] = [\cdots \ell'_k(f_k) \cdots]$
$\delta_k^{(l)} = \sigma'(z_k) \sum_{j \in \text{layer}(l+1)} w_{jk} \delta_j$

Robbins-Monro Algorithm for SGD:

Goal: $\min_\theta \mathbb{E}_Z[f(Z;\theta)] \approx \frac{1}{n}\sum_i \mathcal{L}(y_i, \text{NN}_\theta(x_i))$
Input: learn. rate $\eta(k)$, samples $\mathbf{z}_1, \mathbf{z}_2, \dots \sim Z$
Iteratively: $\theta^{(k)} \leftarrow \theta^{(k-1)} - \eta(k) f(\mathbf{z}_k; \theta^{(k-1)})$
for SGD: $f(z,\theta) = \nabla_\theta \mathcal{L}(y, \text{NN}_\theta(\mathbf{x}))$

Convergence: if $\mathbb{E}_Z[f(\mathbf{z},\theta)]$ satisfies some regulatory conditions and $\eta(k)$ c.f. section **??**.

8.1 Variational Autoencoders

Def: $\mathcal{Z} \xrightarrow[\text{prior}]{p_{\theta'}(z)} \mathcal{Z} \xrightarrow[\text{likelihood}]{\text{dec}_\theta(z) = p_\theta(x|z)} \mathcal{X} \xrightarrow[\text{approx. posterior}]{\text{enc}_\phi(x) = q_\phi(z|x)} \mathcal{Z}$

sample/obs. $\mathcal{X}_i$ from latent representation $\mathcal{Z}$

Train: $\max_{\theta',\theta,\phi} \sum_i \underbrace{\log p_{\theta',\theta}(x_i)}_{(*)\text{ indep. of } Z}$

$$(*) = \mathbb{E}_{Z \sim q_\phi(\cdot|x_i)}\left[\log\left(\frac{p_{\theta',\theta}(x_i, Z)}{p_{\theta',\theta}(Z|x_i)} \frac{q_\phi(Z|x_i)}{q_\phi(Z|x_i)}\right)\right]$$

$$= \overbrace{\mathbb{E}[\log p_\theta(x_i \mid Z)] - D_{\text{KL}}\left(q_\phi(\cdot \mid x_i) \,\|\, p_{\theta'}(\cdot)\right)}^{\mathcal{L}(x_i,\theta,\phi) \equiv \text{ELBO} = \text{Infomax} - \text{Regularisation term}}$$

$$+ D_{\text{KL}}\left(q_\phi(\cdot \mid x_i) \,\|\, p_{\theta',\theta}(\cdot \mid x_i)\right) \geq \mathcal{L}(x_i, \theta, \phi)$$

Train: $\theta^*, \phi^* = \arg\max_{\theta,\phi} \mathcal{L}(x_i, \theta, \phi)$

Requirements for good representation:

- **informative:** $\theta^* = \arg\max_\theta I(X;Z)$
  $= \arg\max_\theta \mathbb{E}_{X,Z}[\log p(X \mid Z)] - const_{\text{w.r.t. } \theta}$
  $\approx \arg\max_\theta \sum_i \mathbb{E}_{Z|X}[\log p(x_i \mid Z)]$
- **disentangled:** components in $\mathcal{Z}$ associated with distinct feature in $\mathcal{X}$ (see D_{KL} in ELBO).
- **robust:** noise in $\mathcal{Z}$ doesn't substantially affect $\mathcal{X}$ (and vice versa). $\to$ choice of approx. post.!

9 Model Selection

Derive posteriors $p^{(i)}(\theta \mid X')$ and $p^{(i)}(\theta \mid X'')$.

ERM: *linear* in noise fluctuations
$p^*(\cdot \mid \cdot) = \arg\min_i \mathbb{E}_{\theta|X'}[-\log p^{(i)}(\theta \mid X'')]$

PA: (only) *quadratic* in noise fluctuations
$p^*(\cdot \mid \cdot) = \arg\max_i \mathbb{E}_{\theta|X'}[p^{(i)}(\theta \mid X'')]$

Note: $\min_p \mathbb{E}_{\theta|X'}[-\log p(\theta \mid X'')] \overset{\text{Jensen}}{\geq} -\max_p \log \mathbb{E}_{\theta|X'}[p(\theta \mid X'')]$

10 Clustering

***k*-means:** $\arg\min_\theta \sum_{i \leq n} \|\mathbf{x}_i - \theta_{c(\mathbf{x}_i)}\|^2$

10.1 Mixture Models

Assume: $\mathbf{x} \sim p(\mathbf{x} \mid \pi_{1\dots k}, \theta_{1\dots k}) = \sum_{c \leq k} \pi_c p(\mathbf{x} \mid \boldsymbol{\theta}_c)$

Find: $\hat{\theta} = \arg\max_\theta p(\mathcal{X} \mid \pi, \theta) = \prod_x p(\mathbf{x} \mid \pi, \theta)$

Gaussian Mixtures: $\to$ $p(\mathbf{x} \mid \theta_c) = p(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma)$

Introduce *latent indicator variables* for mode assignments $M_{xc} \in \{0,1\}$. Then, the **log-likelihood:**

$L(\mathcal{X}, \boldsymbol{M} \mid \boldsymbol{\theta}) = \sum_x \sum_{c \leq k} M_{xc} \log(\pi_c p(\mathbf{x} \mid \boldsymbol{\theta}_c))$

10.1.1 EM-Algorithm for Gaussian Mixtures

E-step: Calculate

$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\boldsymbol{M}|\mathcal{X},\boldsymbol{\theta}^{(t)}}[L(\mathcal{X}, \boldsymbol{M} \mid \boldsymbol{\theta})] = \dots$
$= \sum_x \sum_{c \leq k} \left(\underbrace{\mathbb{E}_{\boldsymbol{M}|\mathcal{X},\boldsymbol{\theta}^{(t)}}[M_{xc}]}\cdot \log \pi_c p(\mathbf{x} \mid \boldsymbol{\theta}_c)\right)$

where $\gamma_{xc} = \frac{p(\mathbf{x}|c,\boldsymbol{\theta}^{(t)})\, p(c|\boldsymbol{\theta}^{(t)})}{p(\mathbf{x}|\boldsymbol{\theta}^{(t)})}$, $\sum_{c \leq k} \gamma_{xc} = 1$

M-step: $\boldsymbol{\theta}^{(t+1)} \in \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$
s.t. $\sum_c \pi_c = 1$. Solve via Lagrangian, yields

$$\pi_c = \frac{1}{|\mathcal{X}|}\sum_x \gamma_{xc}, \quad \boldsymbol{\mu}_c = \frac{\sum_x \gamma_{xc}\mathbf{x}}{\sum_x \gamma_{xc}}, \quad \sigma_c^2 = \frac{\sum_x \gamma_{xc}(\mathbf{x} - \boldsymbol{\mu}_c)^2}{\sum_x \gamma_{xc}}$$

10.2 Non-parametric Bayesian Methods

$\text{Dir}(\mathbf{x} \mid \boldsymbol{\alpha}) = \frac{1}{B(\alpha)} \prod_{k=1}^n x_k^{\alpha_k - 1}$, $B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^n \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^n \alpha_k)}$

Rewrite **Finite mixture models:**
$p(x) = \sum_{k=1}^K \pi_k p(x \mid \theta_k) = \int p(x \mid \theta) G(\theta) \, d\theta$
where $G(\theta) = \sum_{k=1}^\infty \pi_k \delta_{\theta_k}(\theta) \leftarrow$ discrete distr.

Stick-breaking process:
Draw $\theta_k \sim H$ and $\beta_k \sim \text{Beta}(1, \alpha)$ for $k = 1, 2, \dots$
$\pi_k = \beta_k(1 - \sum_{k=1}^{k-1} \pi_i) \implies \pi = \{\pi_k\}_{k=1}^\infty \sim \text{GEM}(\alpha)$
$\implies \sum_{k=1}^\infty \pi_k \delta_{\theta_k}(\theta) = G(\theta) \sim \text{DP}(\alpha, H)$

Sample $\theta^{(1)}, \theta^{(2)}, \dots$ from G. Denote $\theta^{(i)} = \theta_{k_i}$.
$\implies \theta^{(i)}, \theta^{(j)}$ with $k_i = k_j$ belong to same "cluster"

Chinese Rest. Process:
$P(\text{cust}_{n+1} \text{ joins table } \tau \mid \mathcal{P}) = \begin{cases} \frac{|\tau|}{\alpha+n} & \text{if } \tau \in \mathcal{P}, \\ \frac{\alpha}{\alpha+n} & \text{new table} \end{cases}$

$P(\text{partition } \mathcal{P}) = \frac{\alpha^{|\mathcal{P}|}}{\alpha^{(n)}} \prod_{\tau \in \mathcal{P}} (|\tau| - 1)!$

expec. #clusters: $\mathbb{E}[\mathbf{1}] = \sum_{i \leq N} \frac{\alpha}{\alpha + i} \sim \mathcal{O}(\alpha \log N)$

De Finetti: $(X_1, \dots, X_n)$ are inftly **exchangable** RVs *if* $P(X_1, \dots, X_n) = \int \left(\prod_{i=1}^n p(X_i \mid G)\right) dP(G)$

10.2.1 Finite GMM

1. Cluster centers: $\mu_k \sim \mathcal{N}(\mu_0, \sigma_0)$
2. Prob's of clusters: $\pi_{1\dots K} \sim \text{Dir}(\alpha_{1\dots K})$
3. Cluster assignments: $z_i \sim \text{Categorical}(\pi_{1\dots K})$
4. Coordinates of data: $x_i \sim \mathcal{N}(\mu_{z_i}, \sigma_{z_i})$

10.2.2 DP Mixture Model (DP-GMM)

1. Cluster centers: $\mu_k \sim \mathcal{N}(\mu_0, \sigma_0)$, $k = 1, 2, \dots$
2. Prob's of clusters: $\pi = (\pi_1, \pi_2, \dots) \sim \text{GEM}(\alpha)$
3. Cluster assignments: $z_i \sim \text{Categorical}(\pi)$
4. Coordinates of data: $x_i \sim \mathcal{N}(\mu_{z_i}, \sigma)$, $i = 1 \dots N$

Fitting a DP-MM: Collapsed Gibbs sampler
$p(z_i = k \mid \mathbf{z}_{-i}, \mathbf{x}, \alpha, \boldsymbol{\mu}) \propto \text{Prior} \times \text{Likelihood}$
$\propto \begin{cases} \frac{|\mathbf{x}_{-i,k}|}{\alpha+N-1} p(x_i | \mathbf{x}_{-i,k}, \boldsymbol{\mu}) & \text{for existing } k \\ \frac{\alpha}{\alpha+N-1} p(x_i | \boldsymbol{\mu}) & \text{otw.} \end{cases}$

$\mathbf{x}_{-i,c} := \{x_j \mid z_j = c, j \neq i\}$ data assigned to clust. c

11 PAC Learning

Want: Distribution indep. error guarantees!

Expec./Gener. error: $\mathcal{R}(\hat{c}_n) = \boldsymbol{P}_{X,Y}(\hat{c}_n(x) \neq c(x))$

Empirical error: $\hat{\mathcal{R}}_n(\hat{c}_n) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{\hat{c}_n(x) \neq y_i\}}$

PAC learnable: $\mathcal{A}$ can learn a concept class $\mathcal{C}$ from $\mathcal{H}$ if, given a suff. large sample, it outputs a hypothesis that generalizes well w/ high prob.

(1) $0 < \epsilon < 1/2$, $0 < \delta < 1/2$, (2) $\boldsymbol{P}_{X,Y}$ on $\mathcal{X} \times \{0,1\}$:
If $n \geq poly(1/\epsilon, 1/\delta, dim(\mathcal{X}))$,
Then $\boldsymbol{P}_{X,Y}\left(\mathcal{R}(\hat{c}_n) - \inf_{c \in \mathcal{C}} \mathcal{R}(c) \leq \epsilon\right) \geq 1 - \delta$.

If $\mathcal{A}$ runs in time polynomial in $1/\epsilon$ and $1/\delta$, we say that $\mathcal{C}$ is **efficiently PAC learnable**.

11.1 VC Inequality $\quad P(\dots \geq \epsilon) \leq \dots \leq \delta$

Select ERM: $\hat{c}_n^* = \arg\min_{c \in \mathcal{C}} \hat{\mathcal{R}}_n(c)$
Under uniform convergence:

$$\boldsymbol{P}\left(\mathcal{R}(\hat{c}_n^*) - \inf_{c \in \mathcal{C}} \mathcal{R}(c) > \epsilon\right) \leq \boldsymbol{P}\left(\sup_{c \in \mathcal{C}} |\hat{\mathcal{R}}_n(c) - \mathcal{R}(c)| > \frac{\epsilon}{2}\right)$$

- $|\mathcal{C}|$ **Finite:** $\boldsymbol{P}(\sup |\cdots| > \epsilon) \leq 2|\mathcal{C}|\exp(-2n\epsilon^2)$
- $|\mathcal{C}|$ **Unbounded:** $\boldsymbol{P}(\cdots) \leq 9n^{\text{VC}_{\mathcal{C}}} \exp(-\frac{n\epsilon^2}{32})$

11.2 Rectangle Learning

$\boldsymbol{P}((\hat{c}_n^* > \epsilon) \leq |\mathcal{C}| \cdot (1 - \epsilon)^n \leq |\mathcal{C}| \cdot \exp(-n\epsilon) < \delta$

Union bound: $\boldsymbol{P}(\bigcup_i T_i) \leq \sum_i \boldsymbol{P}(T_i)$

12 Appendix

Complete the square: If $p(\mathbf{x}) \propto \exp(-\frac{1}{2}\mathbf{x}^\top \boldsymbol{A}\mathbf{x} + \mathbf{x}^\top \mathbf{b})$, then $p(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{A}^{-1}\mathbf{b}, \boldsymbol{A}^{-1})$

Constrained optimisation:
primal: $\min_x f(\mathbf{x})$ s.t. $g_i(\mathbf{x}) = 0$; $h_j(\mathbf{x}) \leq 0$
Lagrangian: with each $\alpha_j \geq 0$
$\mathcal{L}(\mathbf{x}, \lambda, \alpha) = f(\mathbf{x}) + \sum_i \lambda_i g_i(\mathbf{x}) + \sum_j \alpha_j h_j(\mathbf{x})$
Solve: $\frac{\partial \mathcal{L}}{\partial \mathbf{x}} = 0$; $g_i(\mathbf{x}) = 0$; $\alpha_j \geq 0$; $h_j(\mathbf{x}) \leq 0$
If **Slater's cond.** holds, $\exists \mathbf{x} : g_i(\mathbf{x}) = 0, h_j(\mathbf{x}) < 0$, then we can solve the *dual* instead:
$\max_{\lambda,\alpha}\{\min_x \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\alpha})\}$ s.t. $\alpha_j \geq 0$
Solve: $\frac{\partial \mathcal{L}}{\partial \mathbf{x}} = 0$; $\frac{\partial \mathcal{L}}{\partial \lambda} = 0$; $\alpha_j h_j(\mathbf{x}) = 0$; $\alpha_j \geq 0$

Metrics: $acc = \frac{\text{TP+TN}}{n}$ $prec = \frac{\text{TP}}{\text{TP+FP}}$ $FPR = \frac{\text{FP}}{\text{FP+TN}}$
$Recall/TPR = \frac{\text{TP}}{\text{TP+FN}}$ $balanced\ acc = \frac{1}{n}\sum_i TPR_i$
$F1-score = \frac{2\text{TP}}{2\text{TP+FP+FN}}$ $ROC = FPR/TPR$

Conditional Gaussians:
$P_{X,Y} = \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}\right)$, Σ_{ij} p.s.d.
$\implies Y|X \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma})$, where $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \boldsymbol{\mu}_X)$, $\tilde{\Sigma} = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$