1 Basics

Gaussian

$f(x) = \frac{1}{\sqrt{(2\pi)^d \det\Sigma}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}, \quad \mathcal{N}(x|\mu,\Sigma)$

$X\sim\mathcal{N}(\mu,\Sigma),\ Y=A+BX \Rightarrow Y\sim\mathcal{N}(A+B\mu, B\Sigma B^T)$

Conditionate Gaussians

$\begin{bmatrix}\beta\\x\end{bmatrix} \sim N\left(\begin{bmatrix}\overline{\beta}\\\overline{x}\end{bmatrix},\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\\Sigma_{21} & \Sigma_{22}\end{bmatrix}\right) \Rightarrow \beta|x = y \sim N\left(\overline{\beta}+\Sigma_{12}\Sigma_{22}^{-1}(y-\overline{x}), \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$

Primal Dual problem

Let $\mathcal{P} = \begin{cases}\min_w f(w)\\ g_i(w)=0\ \forall i\\ h_j(w)\leq 0\ \forall j\end{cases}$

Then the Slater's conditions are:
$\exists w\ |g_i(w)=0, h_j(w)<0\ \forall i,j$

The lagrangian is:
$\mathcal{L}(w,\lambda,\alpha) = f(w)+\sum_i\lambda_i g_i(w)+\sum_j\alpha_j h_j(w)$

$\mathcal{D} = \begin{cases}\max_{\lambda,\alpha}\theta(\alpha,\lambda)\\ \theta(\alpha,\lambda)=\min_w\mathcal{L}(w,\lambda,\alpha)\\ \alpha_j(w)\geq 0\ \forall j\end{cases}$

In general the solution of the $\mathcal{D}$ is smaller then $\mathcal{P}$. But if the Slater conditions holds then they are equal. And we get the complementary slackness: $\alpha_j^* h_j(w^*)=0\ \forall$

The optimal $w^* = min_w\mathcal{L}(w,\lambda^*,\alpha^*)$

Calculus

- $\frac{\partial}{\partial\mathbf{x}}(\mathbf{b}^\top\mathbf{x}) = \frac{\partial}{\partial\mathbf{x}}(\mathbf{x}^\top\mathbf{b}) = \mathbf{b}$ • $\frac{\partial}{\partial\mathbf{x}}(\mathbf{x}^\top\mathbf{x}) = 2\mathbf{x}$
- $\frac{\partial}{\partial\mathbf{x}}(\mathbf{x}^\top\mathbf{A}\mathbf{x}) = (\mathbf{A}^\top+\mathbf{A})\mathbf{x} \overset{\text{A sym.}}{=} 2\mathbf{A}\mathbf{x}$
- $\frac{\partial}{\partial\mathbf{x}}(\mathbf{b}^\top\mathbf{A}\mathbf{x}) = \mathbf{A}^\top\mathbf{b}$ • $\frac{\partial}{\partial\mathbf{X}}(\mathbf{c}^\top\mathbf{X}\mathbf{b}) = \mathbf{c}\mathbf{b}^\top$
- $\frac{\partial}{\partial\mathbf{X}}(\mathbf{c}^\top\mathbf{X}^\top\mathbf{b}) = \mathbf{b}\mathbf{c}^\top$ • $\frac{\partial}{\partial\mathbf{x}}(\|\mathbf{x}-\mathbf{b}\|_2) = \frac{\mathbf{x}-\mathbf{b}}{\|\mathbf{x}-\mathbf{b}\|_2}$
- $\frac{\partial}{\partial\mathbf{x}}(\|\mathbf{x}\|_2^2) = \frac{\partial}{\partial\mathbf{x}}(\|\mathbf{x}^\top\mathbf{x}\|_2) = 2\mathbf{x}$ • $\frac{\partial}{\partial\mathbf{X}}(\|\mathbf{X}\|_F^2) = 2\mathbf{X}$
- $x^TAx = Tr(x^TAx) = Tr(xx^TA) = Tr(Axx^T)$
- $\frac{\partial}{\partial A}Tr(AB) = B^T$ • $\frac{\partial}{\partial A}log|A| = A^{-T}$
- $\sigma(x) = \frac{1}{1+e^{-x}}$
- $\nabla\sigma(x) = \sigma(x)(1-\sigma(x)) = \sigma(x)\sigma(-x)$
- $\nabla\tanh(x) = 1-\tanh^2(x)$ • $tanh x = \frac{sinh x}{cosh x} = \frac{e^x-e^{-x}}{e^x+e^x}$

Probability / Statistics

Bayes' Rule $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \frac{P(X|Y)P(Y)}{\sum_{i=1}^{k}P(X|Y_i)P(Y_i)}$

MGF $\mathbf{M}_X(t) = \mathbb{E}[e^{\mathbf{t}^\top\mathbf{X}}]$, $\mathbf{X} = (X_1,..,X_n)$

Markov ineq: $P\{X\geq\epsilon\}\leq\frac{\mathbb{E}[X]}{\epsilon}$ (for nonneg. X)

Boole's inequality: $P(\bigcup_i A_i)\leq\sum_i P(A_i)$

Hoeffding's lemma: $\mathbb{E}[e^{sX}]\leq exp(\frac{1}{8}s^2(b-a)^2)$ where $\mathbb{E}[X]=0$, $P(X\in[a,b])=1$

Hoeffding's: $P\{S_n-\mathbb{E}[S_n]\geq t\}\leq exp(-\frac{2t^2}{\sum_i(b_i-a_i)^2})$

Normalized: $P\{\widetilde{S}_n-\mathbb{E}[\widetilde{S}_n]\geq\epsilon\}\leq exp(-\frac{2n^2\epsilon^2}{\sum_i(b_i-a_i)^2})$

Error bound:
$P\{\sup_{c\in\mathcal{C}}|\hat{\mathcal{R}}_n(c)-\mathcal{R}(c)|>\epsilon\}\leq 2|\mathcal{C}|exp(-2n\epsilon^2)$

Jensen's inequality

X:random variable & φ:convex function $\to$ $\varphi(\mathbb{E}[X])\leq\mathbb{E}[\varphi(X)]$

2 Gaussian Processes

$f\sim GP(\mu,k)\Rightarrow\forall\{x_1,\dots,x_n\}\ \forall n<\infty$
$[f(x_1)\dots f(x_n)]\sim N([\mu(x_1)\dots\mu(x_n)],K)$
where $K_{ij}=k(x_i,x_j)$

Gaussian Process Regression

$f\sim GP(\mu,k)$ then: $f|y_{1:n},x_{1:n}\sim GP(\tilde{\mu},\tilde{k})$

$\tilde{\mu}(z) = \mu(z)+K_{D,z}^T(K_{DD}+\epsilon I_n)^{-1}(y_{1:n}-\mu(x_{1:n}))$

$\tilde{k}(z_1,z_2) = k(z_1,z_2)-K_{D,z_1}^T(K_{DD}+\epsilon I_n)^{-1}K_{D,z_2}$

Where: $K_{D,z} = [k(x_1,z)\dots k(x_n,z)]^T$
$[K_{DD}]_{ij} = k(x_i,x_j)$

2.1 Kernels

$k(x,y)$ is a kernel if it's symmetric semidefinite positive:
$\forall\{x_1,\dots,x_n\}$ then for the Gram Matrix $[K]_{ij}=k(x_i,x_j)$ holds $c^TKc\geq 0\forall c$

Closure Properties: psd prop. closed under pointwise limits (since each K_n is a kernel)

$k(x,y) = k_1(x,y)+k_2(x,y)$, $k(x,y) = k_1(x,y)k_2(x,y)$
$k(x,y) = f(x)f(y)$, $k(x,y) = k_3(\phi(x),\phi(y))$
$k(x,y) = \exp(\alpha k_1(x,y)), \alpha>0$, $|X\cap Y| = kernel$
$k(x,y) = p(k_1(x,y))$, $p(\cdot)$ polynomial with pos. coeff.
$k(x,y) = k_1(x,y)/\sqrt{(k_1(x,x)k_1(y,y))}$

Gaussian (rbf): $k(x,y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2})$ inf.dim.

Sigmoid: $k(x,y) = \tanh(k\cdot x^Ty-b)$ not valid for $\forall k,b$

Polynomial: $k(x,y) = (x^Ty+c)^d, d\in N, c\geq 0$

Periodic: $k(x,y) = \sigma^2 exp(\frac{2\sin^2(\pi|x-y|/p)}{\ell^2})$

3 Statistics Recap

Estimation

Consistency: $\hat{\theta}_n\overset{P}{\to}\theta$, i.e. $\forall\epsilon P\{|\hat{\theta}_n-\theta|\geq\epsilon\}\overset{n\to\infty}{\longrightarrow}0$

Asymptotic normality: $\sqrt{N}(\theta-\hat{\theta}_n)\to\mathcal{N}(0,J^{-1}IJ^{-1})$

Asymptotic efficiency: $\hat{\theta}_n$ reaches the Rar Cramer bound in the limit, i.e. $\lim_{n\to\infty}(V[\hat{\theta}_n]\mathcal{I}_n(\theta))^{-1}=1$

Rao-Cramer

$\Lambda = \frac{\partial\log\mathbb{P}(x|\theta)}{\partial\theta}$ (score function), $E[\Lambda]=0$

Fisher information: $\mathcal{I}(\theta) = \mathbb{V}[\Lambda]$

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

If the model is realizable then $\mathcal{I}=\mathcal{J}$

Oss: For the whole model:
$\mathcal{I}_n = \mathbb{V}\left[\frac{\partial\log\mathbb{P}(x_i,i=1:n|\theta)}{\partial\theta}\right] = n\mathcal{I}$

let $b(\hat{\theta}) = \mathbb{E}\left[\hat{\theta}\right]-\theta$

MSE bound: $E[(\hat{\theta}-\theta)^2]\geq\frac{[1+b'(\hat{\theta})]^2}{nE[\Lambda^2]}+b(\hat{\theta})^2$

Biased estimators: $var(\hat{\theta})\geq\frac{[1+b'(\hat{\theta})]^2}{n\mathcal{I}(\theta)}$

Efficiency: $e(\hat{\theta}) = \frac{I(\theta)^{-1}}{var(\hat{\theta})}\leq 1$

Cauchy-Schwarz: $|E(XY)|^2\leq E(X^2)E(Y^2)$

4 Linear Regression

$y = X\beta+\epsilon$ where $y\in\mathbb{R}^n$, $X\in\mathbb{R}^{n\times d}$, $\beta\in\mathbb{R}^d$

Risk Decomposition Theorem

$\mathbb{E}_{Y,D}\left[\left(Y-\hat{f}(x_0)\right)^2\right] = Bias+Vairance+Noise$

$Bias = \left(\mathbb{E}[Y|X=x_0]-\mathbb{E}_D\left[\hat{f}(x_0)\right]\right)^2$

$Variance = \mathbb{E}_D\left[\left(\mathbb{E}_D\left[\hat{f}(x_0)\right]-\hat{f}(x_0)\right)^2\right]$

$Noise = \mathbb{E}_Y\left[(Y-\mathbb{E}[Y|X=x_0])^2\right]$

Combination of Regression Models:

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

$\mathbb{V}[\hat{f}(x)] = \frac{1}{B^2}\sum_i\mathbb{V}[\hat{f}_i(x)]+\frac{1}{B^2}\sum_{i\neq j}cov[\hat{f}_i(x),\hat{f}_j(x)]\approx\frac{\sigma^2}{B}$

Minimum square linear regression

$\hat{\beta} = \arg\min_\beta\|X\beta-y\|\Rightarrow\hat{\beta} = \left(X^TX\right)^{-1}X^Ty$.

Here $\hat{\beta}$ is the BLUE (Best Linear Unbiased Estimator)

Lasso regression

$\hat{\beta} = \arg\min_\beta\|X\beta-y\|+\lambda\|\beta\|_1\Rightarrow\hat{\beta}$ = No closed form (LARS algorithm) but it is a convex problem

Bayesian prior: $p(\beta_i) = \frac{1}{4\sigma^2}exp\left(-|\beta_i|\frac{\lambda}{2\sigma^2}\right)$

Const. opt. $\hat{\beta} = \arg\min_\beta\|X\beta-y\|$ s.t. $\|\beta\|_1<s_\lambda$

Ridge regression

$\hat{\beta} = \arg\min_\beta\|X\beta-y\|+\lambda\|\beta\|_2\Rightarrow\hat{\beta} = \left(X^TX+\lambda I\right)^{-1}X^Ty$

Bayesian prior $p(\beta) = N(0,\frac{\sigma^2}{\lambda}I)$

Oss: if instead $p(\beta) = N(0,\Lambda^{-1})$ then $\hat{\beta} = \left(X^TX+\sigma^2\Lambda\right)^{-1}X^Ty$

Const. opt. $\hat{\beta} = \arg\min_\beta\|X\beta-y\|$ s.t. $\|\beta\|_2<s_\lambda$

Let μ_i be the singular values of X then $|\left(X^TX\right)^{-1}X^T| = \prod^i\frac{1}{\mu_i}$. And $|\left(X^TX+\lambda I\right)^{-1}X^T| = \prod^i\frac{\mu_i^2}{\mu_i^2+\lambda}$. Therefore if $\mu_i\simeq 0$ with Ridge we have no problems (stable results against inter column linear dependence)

5 Numerical Estimating Methods

Actual Risk: $\mathcal{R}(f) := \mathbb{E}_{x,y}[(y-f(x))^2]$

Empiricial Risk: $\hat{\mathcal{R}}(f) = \frac{1}{n}\sum_i(y_i-f(x_i))^2$

Generalization Error: $G(f) = |\hat{\mathcal{R}}(f)-\mathcal{R}(f)|$

K-fold cross validation

$\hat{f}^{-\nu}\in\arg\min_f\frac{1}{|Z^{-\nu}|}\sum_{i\in Z^{-\nu}}(y_i-f(x_i))^2$

$\hat{\mathcal{R}}^{cv} = \frac{1}{n}\sum_i(y_i-\hat{f}^{-\kappa(i)}(x_i))^2$, $k(i)$ is fold i^{th} fold

Problem: systematic tendency to underfit.

Leave-one-out (LOOCV) = K-fold $(K=n)$

Jackknife (Estiamte the bias of estiamtor)

$\text{bias}^{JK} = (n-1)(\hat{\theta}-\tilde{\theta})$ with $\tilde{\theta} = \frac{1}{n}\sum_{i=1}^n\hat{\theta}^{(-i)}$ and $\hat{\theta}^{(-i)}$is the leave out i estiamtor.

The corrected estimator is: $\hat{\theta}^{JK} = \hat{\theta}-\text{bias}^{JK}$

Information Criteria

$BIC = ln(n)k-2ln(\hat{L})$, $AIC = 2k-2ln(\hat{L})$

$TIC = 2trace[I_1(\theta_k)J_1^{-1}(\theta_k)]-2ln(\hat{L})$, where k: num. params, n: num. data points, likelihood: $\hat{L} = p(X|\theta_k,M)$

6 Classification

Loss-Functions

True class: $y\in\{-1,1\}$, pred. $z\in[-1,1]$

Cross-entropy (log loss): ($y' = \frac{(1+y)}{2}$ and $z' = \frac{(1+z)}{2}$) $L(y',z') = -[y'log(z')+(1-y')log(1-z')]$

Hinge Loss: $L(y,z) = max(0,1-yz)$

Perceptron Loss: $L(y,z) = max(0,-yz)$

Logistic loss: $L(y,z) = log(1+exp(-yz))$

Square loss: $L(y,z) = \frac{1}{2}(1-yz)^2$

Exponential loss: $L(y,z) = exp(-yz)$

Binomial deviance: $L(y,z) = 1+exp(-2yz)$

0/1 Loss: $L(y,z) = \mathbb{I}\{sign(z)\neq y\}$

Probabilistic generative approach

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

where $p(y|x)$ is found from $p(y,x)$ which is itself estimated somehow

Probabilistic discriminative approach

Like Probabilistic generative approach but we estimate $p(y|x)$ directly.

$p(y|x) = \arg\max_w\mathcal{L}(\mathcal{Z}_{train},w) = \arg\max_w\sum_i\log p(y_i|x_i,w)$ where $p(y|x;w) = \sigma(w^Tx+w_0)$. We can gradient descent on $-\mathcal{L}$

Discriminative approach

Directly look for:

$c^* = \arg\min_c\hat{\mathcal{R}}(c,\mathcal{Z}_{train}) = \arg\min_c\frac{1}{n}\sum_{i=1}^n L(y_i,c(x_i))$

Percepton Algo

Find w, w_0 s.t. $y_i w^t x_i > 0\ \forall i$. Gradient descent on $L(y, c(x)) = -y w^T x \mathbb{I}_{(-\inf,0)}\left(y w^T x\right)$
or $L(y, c(x)) = \min_{\alpha_{1:n}} \sum_{i=1}^n \max[0, -\sum_{j=1}^n \alpha_j y_i y_j x_i^T x_j]$

Fischer Discriminant

$w^* = \arg\max_w \frac{w^T S_B w}{w^T S_w w} = S_w^{-1}(\overline{x}_0 - \overline{x}_1)$ where:
$S_B = (\overline{x}_0 - \overline{x}_1)^T (\overline{x}_0 - \overline{x}_1)$
$S_w = \hat{Cov}(C_0) + \hat{Cov}(C_1)$ Sample variance matrixes for each cluster.
Fit a mixture of gaussians on $w^{*T} x$ insted of x

7 SVM

Like Percepton but maximizing the margin. Equivalent to

$\mathcal{P} = \begin{cases} \min_{w,w_0} \frac{\|w\|^2}{2} \\ y_i(w^T x_i + w_0) \geq 1\ \forall i \end{cases}$ where the margin size is $\frac{2}{\|w\|^2}$.

X^+, X^- are separable $\Rightarrow$ Slater conditions $\Rightarrow$
$\mathcal{D} = \begin{cases} \max_\alpha \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ \alpha_i \geq 0\ \forall i \end{cases}$

Complementary slackness $\alpha_i^* h_i(w^*) = 0$ so either $\alpha_i^* = 0$ or x_i is a Support Vector

Soft margin SVM

We add a C parameter (C small $\Rightarrow$ soft):

$\mathcal{P} = \begin{cases} \min_{w,w_0,\xi} \frac{\|w\|^2}{2} + C \sum_i \xi_i \\ y_i(w^T x_i + w_0) \geq 1 - \xi_i\ \forall i \\ \xi_i \geq 0\ \forall i \end{cases}$

$\mathcal{D} = \begin{cases} \max_{\alpha_i} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ 0 \leq \alpha_i \leq C\ \forall i \end{cases}$

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

$y = sgn\left(w^{*T} x\right) = sgn\left(\left(\sum_i \alpha_i^* y_i x_i\right)^T x_j\right)$

Non linear SVM: $x_i^T x_j \rightarrow \phi(x_i)^T \phi(x_j) \rightarrow k(x_i, x_j)$

Multiclass SVM (ovr)

Train a binary classifier for each class (one vs the rest). Then I assign a score $f_c(x) = w_c^T x$.
Predicitons: $c^* = \arg\max_c f_c(x)$

Structured SVM

Too many class for ovr. $\Psi : X \times Y \rightarrow \mathbb{R}^{m+d}$ is called Joint feature map $\mathcal{P} =$
$\begin{cases} \min_{w,w_0} \frac{\|w\|^2}{2} + \frac{C}{n} \sum_{i=1}^n \xi_i \\ w^T \Psi(x_i, y_i) \geq \Delta(y_i, y') + w^T \Psi(x_i, y') - \xi_i\ \forall i\ \forall y' \neq y_i \\ \xi \geq 0\ \forall i \end{cases}$

Theorem Δ as Loss (Structured SVM in Statistical Learning):
$\hat{\mathcal{R}}(\mathcal{Z}_{train}) \doteq \frac{1}{n} \sum_{i=1}^n \Delta(y_i, c_{w^*}(x_i)) \leq \frac{1}{n} \sum_{i=1}^n \xi_i^*$

8 Ensemble method

Bagging

We train $b^{(1)}, \ldots, b^{(M)}$ different classifiers.
Then $\overline{b}(x) = \begin{cases} \frac{1}{M} \sum_{i=1}^n b^{(i)}(x) & \text{regression} \\ \text{majority}\left(b^{(i)}\right) & \text{classification} \end{cases}$

Works: Covariance small (different subset for training or bootstrap), Variance small (similar behaviour of weak learners), biases weakly affected.
Bias↓&Var.↓: Use complex decision tree (bias↓), ensemble mult. decision trees (var↓)

Random Forest

Is a sort of bagging with decision trees. At each splitting node we draw m features and we pick the splitting one only among them (↓ correletion among trees). We also use Bootstrap

Adaboost

Boosting: Train weak learners sequentially on all data, but reweight misclassifed samples higher, Bias ↓
Initialize weights $w_i = 1/n$, for b=1:B do:
1. Fit classifier $c_b(x)$ with weights w_i
2. Compute error $\epsilon_b = \sum_i w_i^{(b)} \mathbb{1}_{[c_b(x_i) \neq y_i]} / \sum_i w_i^{(b)}$
3. Compute coeff. $\alpha_b = log(\frac{1-\epsilon_b}{\epsilon_b})$
4. Update weights $w_i = w_i \exp(\alpha_b \mathbb{1}_{[y_i \neq c_b(x_i)]})$

Return $\hat{c}_B(x) = \text{sign}\left(\sum_{b=1}^B \alpha_b c_b(x)\right)$
Loss: Exponential loss $L(y, y') = exp(-yy')$
Model: Forward Sationary Adaptive.
Oss: Self averaging algos that train Spiky interpolating classifiers.
AdaBoost trains max-margin classifier.

9 Mixtures Models (Unsupervised Learning)

K-means

We find $\mu_1, \ldots, \mu_k$ such that our predictions are $c(x) : \mathbb{R}^d \rightarrow \{1, \ldots, k\}$.
Find $c(\cdot)$ and $\mu_i \forall i$ that minimize:
$\mathcal{R}^{km}(c, \mu_i \forall i) = \sum_x \left\|x - \mu_{c(x)}\right\|^2$

Initialize $\mu_i \forall i$;
while *μ_i are changing* **do**
  $c(x) \leftarrow \arg\min_c \left\|x - \mu_c\right\|^2\ \forall x$;
  $\mu_\alpha = \frac{1}{n_\alpha} \sum_{x:c(x)=\alpha} x\ \forall \alpha$;

Gaussian Mixtures

1) Draw $z \sim \pi$ Categorical.
2) Draw $x \sim N(\mu_z, \Sigma_z)$

Expectation Maximization

Initialize $\theta^0 = \pi^0, \mu^0, \sigma^{2\,0}$;
while $\left\|\theta^{j+1} - \theta^j\right\| > \epsilon$ **do**
  E-step:
  $\gamma_{xc} \doteq \mathbb{E}\left[M_{xc} | X, \theta^j\right] =$
  $\frac{p(X|c,\theta^j), p(c|\theta^j)}{p(x|\theta^j)} = \frac{N(\mu_c^j, \sigma_c^{2\,j})\pi_c^j}{\sum_\nu \pi_\nu N(\mu_\nu, \sigma_\nu^{2\,j})}$
  $Q(\theta, \theta_j) = \mathbb{E}\left[L(X, X_L|\theta)|\theta_j\right] =$
  $\sum_{x \in X} \sum_c (\gamma_{xc} \log(\pi_c P(x|\theta_c)))$
  M-step: $\theta_{j+1} = \arg\max_\theta Q(\theta, \theta_j)$
  $\pi_c^{j+1} = \frac{1}{|X|} \sum_{x \in X} \gamma_{xc}$
  $\mu_c^{j+1} = \frac{\sum_{x \in X} \gamma_{xc} x}{\sum_{x \in X} \gamma_{xc}}$
  $\sigma_c^{2\,j+1} = \frac{\sum_{x \in X} \gamma_{xc}(x - \mu_c)^2}{\sum_{x \in X} \gamma_{xc}}$

Where $M_{xc} = \mathbb{I}_{\{x \text{ generated by } c\}}(x)$

10 Neural Network

Backpropagation

Let $\Phi(x) = f_{\theta_n}^{(n)} \circ f_{\theta_{n-1}}^{(n-1)} \circ \cdots \circ f_{\theta_1}^{(1)}(x)$
$\partial_\Phi f^{(i)} \doteq \partial_z f^{(i)}(z, \theta_i)|_{z=\Phi^{(i-1)}(x)}$
$\partial_\theta f^{(i)} \doteq \partial_z f^{(i)}(\Phi^{(i-1)}(x), \theta)|_{\theta=\theta_i}$

Result: $\partial_{\theta_i} \Phi(x) \forall i$
Initialize $B = 1$;
for $i \leftarrow n, n-1, \ldots, 1$ **do**
  $\partial_{\theta_i} \Phi(x) \leftarrow B \partial_\theta f^{(i)}$;
  $B \leftarrow B \partial_\Phi f^{(i)}$;

Once we have this we can $\nabla \downarrow$

Stocastic Gradient Descent

Result: optimal θ^*
Initialize θ;
while *Test error is decreasing* **do**
  $\nabla_\theta Loss = \sum_{(x,y) \in S_k} \nabla_\theta \mathcal{L}(NN(x), y)$;
  $\theta \leftarrow \theta - \eta(k) \nabla_\theta Loss$;

Oss: $S_k \in D$ and changes at each iteration (Mini Batch)
Oss: As long as $\sum_k \eta(k) = \infty$ and $\sum_k \eta^2(k) < \infty$ the SGD converges
Advantages over Normal Gradient Descent: 1) Can handle large Dataset 2) Faster improvment (with regards to time, not iterations) 3) Escapes local minima 4) Lower generalization error

11 Autoencoders

Infomax principle

Let $I(X, Y) \doteq H(X) - H(X|Y)$ be the mutual information.
$\theta^* = \arg\max_\theta I(X, enc_\theta X)$
$\theta^* \simeq \arg\max_\theta \sum_i \mathbb{E}_Z [\log p(x_i|Z)]$
It is informative but not Disentangled and Robust

Variation Autoencoders

Let $p_{\theta'}(\cdot)$ be our prior, $p_\theta(\cdot|z)$ be our likelihood, $q_\lambda(z|x)$ the postirior.
$\theta^*, \theta'^*, \lambda^* = \arg\max \sum_{i=1}^n \log p_{\theta,\theta'}(x_i)$
In practice we maximize the Evidence Lower Bounds:
$ELBO = \mathbb{E}_{Z \sim q_\lambda(\cdot, x_i)}[\log p_\theta(x_i|z)]$ (infomax) $-KL(q_\lambda(\cdot, x_i) \| p_{\theta'})$ (- distance from the prior)

12 Nonparametric Bayesian methods

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

Chinese Restourant Process

$p(\text{cust}_{n+1} \text{ joins table } \tau | \mathcal{P}) = \begin{cases} \frac{|\tau|}{\alpha + n} & \tau \in \mathcal{P} \\ \frac{\alpha}{\alpha + n} & \tau \notin \mathcal{P} \end{cases}$

de Finetti: $p(X_1, \ldots, X_n) = \int (\prod_{i=1}^n p(x_i|G)) dP(G)$

Gibbs Sampling

DP generative model:
- Centers of the clusters: $\mu_k \sim \mathcal{N}(\mu_0, \sigma_0)$
- Prob.s of clusters: $\rho = \{\rho_k\}_{k=1}^\infty \sim GEM(\alpha)$
- Assignments to clusters: $z_i \sim Categorical(\rho)$
- Coordinates of data points: $\mathcal{N}(\mu_{z_i}, \sigma)$

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

13 PAC Learning

Empirical error: $\hat{\mathcal{R}}_n(c) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{c(x_i) \neq y\}}$
Expected error: $\mathcal{R}(c) = P\{c(x) \neq y\}$
ERM: $\hat{c}_n^* = \arg\min_{c \in \mathcal{C}} \hat{\mathcal{R}}_n(c)$
opt: $c^* \in \min_{c \in \mathcal{C}} \mathcal{R}(c)$, $|\mathcal{C}|$ finite
Generalization error: $\mathcal{R}(\hat{c}_n^*) = P\{\hat{c}_n^*(x) \neq y\}$
$\mathcal{A}$ can learn c if $\exists \pi \in$ Polynomials s.t.:
- $\forall$ distribution $\mathcal{D}$ over X
- $\forall \epsilon \in (1, \frac{1}{2})$, $\forall \delta \in (1, \frac{1}{2})$
- $\forall n \geq \pi(\frac{1}{\epsilon}, \frac{1}{\delta}, \text{size}(c))$

then $\mathbb{P}_{\mathcal{Z} \sim \mathcal{D}}(\mathcal{R}(\mathcal{A}(\mathcal{Z})) \leq \epsilon) \geq 1 - \delta$

VC ineq.: $\mathcal{R}(\hat{c}_n^*) - \inf_{c \in \mathcal{C}} \mathcal{R}(c) \leq 2 \sup_{c \in \mathcal{C}} |\hat{\mathcal{R}}_n(c) - \mathcal{R}(c)|$
$P\{\mathcal{R}(\hat{c}_n^*) - \mathcal{R}(c^*) > \epsilon\} \leq P\{\sup_{c \in \mathcal{C}} |\hat{\mathcal{R}}_n(c) - \mathcal{R}(c)| > \frac{\epsilon}{2}\}$
$\leq 2|\mathcal{C}| exp(-2n\epsilon^2/4)$ if $\mathcal{C}$ is finite
$\leq 9n^{\mathcal{VC}_\mathcal{C}} exp(-n\epsilon^2/32)$
where the $\mathcal{VC}$ dimension of a function class $\mathcal{C}$ is the maximum number of points that can be arranged so that $\mathcal{C}$ shatters them.