Side Notes

- Nomial scale: $f : R \rightarrow R : f$ is bijective ,Ordinal scale: $f : R \rightarrow R : f(x_1) < f(x_2)\ \forall x_1 < x_2$, Interval scale : $f : R \rightarrow R : f(x) = ax + c$ and a is positive , Ration scale: $f : R \rightarrow R : f(x) = ax$ and a is positive , Absolute scale: $f : R \rightarrow R : f$ is identity map
- The Bayes optimal classifier with the maximum a posteriori rule yields **linear classification rules for Gaussian class conditional densities when the variances are the same**. So in this situation linear classifier is optimal because MAP classifier is optimal regarding to 0-1 loss. When variances differ, it leads to a quadratic boundry.
- Early stopping is like putting noise in the dataset.
- WINNOW algorithm's motivation is to use exponential updates instead of additive updates of perceptron for a faster convergence.
- In SSVM the number of features depends on the dimensionality of the joint feature map only and is "independent" of the number of classes.
- Prediction is hard in SSVM because for a fixed w finding the max f across all classes is combinatorial and needs extra assumptions to become possible.
- Arding is adaptive reweighting and combining a extension of bagging.
- $(1 - x)^n \leq e^{-x}$
- VC-dimension of convex polygons in R^2 is ∞. and VC-dimension of convex polygons with at most k vertices is $2k + 1$

Derivatives

$\frac{\partial}{\partial \Sigma^{-1}} log|\Sigma| = -\Sigma$
$\frac{\partial}{\partial \Sigma^{-1}} a^T \Sigma^{-1} a = aa^T$
$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top \mathbf{A}\mathbf{x}) = (\mathbf{A}^\top + \mathbf{A})\mathbf{x}$
$\frac{\partial}{\partial \mathbf{X}}(|\mathbf{X}|) = |\mathbf{X}| \cdot \mathbf{X}^{-1} \qquad |X| = 1/|X^{-1}|$
$\frac{\partial}{\partial x}(\mathbf{Y}^{-1}) = -\mathbf{Y}^{-1}\frac{\partial \mathbf{Y}}{\partial x}\mathbf{Y}^{-1}$

Taylor Expansion

$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \ldots$
$f(x) \approx f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(b)}{2!}(x-a)^2$ and b is in the line between a and x.

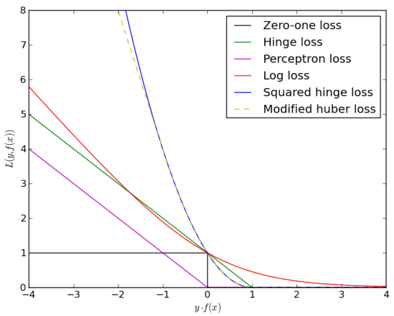
Cauchy Swartz

$E^2[XY] \leq E[X^2]E[Y^2]$

Risk, Bias, Variance

- Conditional expected risk: $R(X, f(X)) = \int_y Loss(y, f(X))P(y|X)dy$
- Total expected risk: $R(f(X)) = \int_x \int_y Loss(y, f(x))p(y|x)dydx$
- How good we are? $P(|\hat{R}(\hat{f}(X^{test}), \mathcal{Z}^{test}) - E_X[R(\hat{f}(X), X)| > \epsilon)$
- Bias: $\mathbb{E}_x[\mathbb{E}_D \hat{f}_D(x) - \mathbb{E}_y[Y|X = x]]$
- Variance: $\mathbb{E}_X[\mathbb{E}_D[(\hat{f}_D(X) - \mathbb{E}_D[\hat{f}_D(X)])^2]]$
- Noise: $\mathbb{E}_Y[Y - \mathbb{E}_y[Y|X = x]]$

Loss functions(Classification)



Bayesian Inference for Guassian

$p(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$
$p(x|\mu) = \mathcal{N}(\mu, \sigma^2)$
$\Rightarrow p(\mu|x) = \mathcal{N}(\mu_N, \sigma_N)$ where $\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML}$ where μ_{ML} is $\frac{1}{N}\sum x_i$ and $\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$

Effective sample size for bayesian learning in guassains is $\frac{\sigma}{\sigma_0}$ prediction on new sample x is: $p(x|\mathcal{X}) = \int p(x|\mu)p(\mu|\mathcal{X})d\mu \sim \mathcal{N}(\mu_N, \sigma^2 + \sigma_N^2)$

Multivariate Gaussian

$\mathbb{P}(\underline{x}) = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} exp(\frac{-1}{2}(\underline{x} - \mu)^T\Sigma^{-1}(\underline{x} - \mu))$ and in this case the posterior is: $\Sigma_n^{-1} = n\Sigma^{-1} + \Sigma_0^{-1}$ and $\Sigma_n^{-1}\mu_n = n\Sigma^{-1}\hat{\mu}_n + \Sigma_0^{-1}\mu_0$

Coefficient Matching for Posterior

$ax^TAX + bx^TB + c \Rightarrow \Sigma = A^{-1},\ \mu = \Sigma B$

$P(\begin{bmatrix}\mathbf{a1}\\ \mathbf{a2}\end{bmatrix}) = \mathcal{N}(\begin{bmatrix}\mathbf{a1}\\ \mathbf{a2}\end{bmatrix} | \begin{bmatrix}\mathbf{u1}\\ \mathbf{u2}\end{bmatrix}, \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix})$

$p(\mathbf{a2}|\mathbf{a1}) = \mathcal{N}(\mathbf{a2}|\mathbf{u2} + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{a1} - \mathbf{u1}), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$

Properties of Estimators

- unbiased: $E[\hat{\theta}_n] = \theta_0$
- consistent: $\forall \epsilon\ P(|\hat{\theta}_n - \theta_0| > \epsilon) \xrightarrow{n \rightarrow \infty} 0$
- asymptotically normal: $\frac{(\hat{\theta}_n - \theta_0)}{\sqrt{N}}$ converges in distribution to a random variable with distribution $\mathcal{N}(0, J^{-1}(\theta)I(\theta)J^{-1}(\theta))$ and $I(\theta) = var[\frac{\partial logP(x|\theta)}{\partial \theta}]$ and $J(\theta) = -E[\frac{\partial^2 logP(x|\theta)}{\partial\theta\partial\theta^T}]$ and **I and J are equal iff true θ which is θ^* is equal to** θ_0
- asymptotically efficient: $E[(\hat{\theta}_n - \theta_0)^2]$ is **minimizes** as $n \rightarrow \infty$
- θ_{ML} **is not** always unbiased. It **is** consistent, asymptotically normal and asymptotically efficient but if N is finite it is not efficient.(Stein)

Rao-Cramer bound

$E[(\hat{\theta}_n - \theta_0)^2] \geq \frac{1}{I_n(\theta_0)}$ for any unbiased estimator $\hat{\theta_n}$ where $I_n(\theta_0)$ is Fisher's information:

$$I_n(\theta_0) = Var[V] = Var[\frac{\partial}{\partial \theta}[logP(y_1, y_2, ..., y_n|\theta)]_{\theta_0}] = E[V^2]$$

For switching integral and derivative, the function shoud be continuously differentiable and integral convergence should be uniform
For biased estimators:

$$E[(\hat{\theta}_n - \theta_0)^2] \geq \frac{1 + \frac{\partial}{\partial\theta}b_{\hat{\theta}}}{I_n(\theta_0)} + b_{\hat{\theta}}^2$$

Convergence in Distribution

random variable $r_1, r_2, ..., r_n$ converges to r in distribution if for every **continuous and bounded** function f we have: $E[f(r_n)] \xrightarrow{n \rightarrow \infty} E[r]$

Linear Regression

Let $f(x, \beta) = x^T\beta$ then for $X^{n \times d}$: $\hat{\beta} = (X^TX)^{-1}X^TY$. If $y = X\beta + \epsilon$ then we know $\hat{\beta} \sim \mathcal{N}(\beta, (X^TX)^{-1}\sigma^2)$. Least square estimate has lowest variance among all linear unbiased estimates e.g. $E_y[\hat{\beta}(y)]$ is minimum.

- **Ridge**: $RSS(\beta) = (X - \beta Y)^T(X - \beta Y) + \lambda\beta^T\beta$ which is equal to $P(Y|\beta, X) \sim \mathcal{N}(x^T\beta, \sigma^2 I)$ and $\beta \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda}I)$
  Bayesian : $p(\beta|\Gamma) \sim \mathcal{N}(0, \Gamma^{-1})$ and $p(\beta|\Gamma, X, y) = \mathcal{N}((X^TX + \sigma^2\gamma)^{-1}X^Ty, \sigma^2(X^TX + \sigma^2\Gamma)^{-1}))$
- **Lasso** : $RSS(\beta) = (X - \beta Y)^T(X - \beta Y) + \lambda|\beta|$ which is equal to $P(Y|\beta, X) \sim \mathcal{N}(x^T\beta, \sigma^2 I)$ and $p(\beta_i) = \frac{\lambda}{4\sigma^2}exp(-|\beta|\frac{\lambda}{2\sigma^2})$
- Ridge regression shrinks directions of column space X which have low variance(low predictive value) the shrinkage factor is $\frac{d_j}{d_j + \lambda}$ where d_j is eigen value j of X

Gaussian Process

The joint distribution is $P(y, f^*) \sim \mathcal{N}(0, \begin{bmatrix}K(X,X) + \sigma^2 I & K(X, X^*)\\ K(X^*, X) & K(X^*, X^*) + \sigma^2 I\end{bmatrix})$ Then $P(f^*|y) \sim \mathcal{N}(\mu_{f^*}, cov(f^*))$ where $\mu_{f^*} = K(X^*, X)[K(X, X) + \sigma^2 I]^{-1}y$ and $cov(f^*) = K(X^*, X^*) + \sigma^2 I - K(X^*, X)[K(X, X) + \sigma^2 I]^{-1}K(X, X^*)$

Validation

Cross-Validation

$\hat{R}^{CV} = \frac{1}{N}\sum_i(y_i - \hat{f}^{-\kappa(i)}(x_i))^2$
If $k \uparrow$ (LOOCV) $\Rightarrow$ bias $\downarrow$, var $\uparrow$, if $k \downarrow$ then bias $\uparrow$, var $\downarrow$. Rule of thumb for K-fold cv? $k = min(\sqrt(n), 10)$.

Bootstrapping

training data $\rightarrow$ surrogate data source. sample $Z^{*1}, ..., Z^{*B}$ from training data with replacement. Then compute mean and variance: mean is $\bar{S} = \frac{1}{B}\sum_{b \leq B} S(Z^{*b})$ and variance is $\sigma^2(S) = \frac{1}{B-1}\sum_{b \leq B}(S(Z^{*b}) - \bar{S})^2$
Consistency of bootstrapping means $lim_{B \rightarrow \infty}\frac{1}{B}\sum_{b \leq B}(S(Z^{*b}) - \bar{S})^2 = V_F[S(Z)]$. How to use it for **model selection**? $R^* = \frac{1}{B}\frac{1}{N}\sum_b\sum_i l(y_i, \hat{f}^{b*}(x_i))$ which is **too optimistic** because each sample i with probability $1 - \frac{1}{e} \approx 0.6$ appears in bootstrap b

Jackknife Bias Estimator

$bias^{JK} = (n - 1)(\tilde{S}_n - \hat{S}_n)$ where $\tilde{S}_n = \frac{1}{n}\sum_i \hat{S}_{n-1}^{-i}$
Then $E[bias^{JK}] = \frac{a_1}{n} + O(n^{-2})$ and the final estimator is $\hat{S}^{JK} = \hat{S}_n - bias^{JK}$. *Bias corrected estimators can have a considerably larger variance than uncorrected estimators*

Example Exercise Jackknife
Setup: $X \sim \mathcal{U}[0, \theta], \hat{S}_n = max(X) = X_n$.
1) Expected value of estimator:
1.1) CDF $P(X_n \leq x) = (\frac{x}{\theta})^n$
1.2) PDF: $p(x) = \frac{\delta P(X_n \leq x)}{\delta x} = n\frac{x^{n-1}}{\theta^n}$
1.3) $\mathbb{E}(\hat{S}_n) = \int_0^\theta xn\frac{x^{n-1}}{\theta^n} = \frac{n}{n+1}\theta$
2) $\hat{S}_{n-1}^{i-1} = X_n$ if $i \neq i^*$ else X_{n-1}
3) $\hat{S}^{JK} = X_n + \frac{n-1}{n}(X_n - X_{n-1})$

4) Expected value of X_{n-1}
4.1) $P(X_{n-1} \le x) = (\frac{x}{\theta})^n + (\frac{x}{\theta})^{n-1}\frac{\theta - x}{\theta}n$
4.2) $p(x) = \frac{n(n-1)}{\theta}(\frac{x}{\theta})^{n-2}(1-\frac{x}{\theta})$
4.3) $\mathbb{E}(X_{n-1}) = \frac{n-1}{n+1}\theta$
5) $\mathbb{E}(\hat{S}_n^{JK}) = (1 - \frac{1}{n^2+n})\theta$

Neyman-Pearson Lemma

$A_n(T) = \{\frac{P_0(x_1,x_2,\dots,x_n)}{P_1(x_1,x_2,\dots,x_n)} \ge T\}$. False positive is $P_0(A^c(T)) = \alpha^*$ and false negative is $P_1(A(T))$ then it holds $\beta \ge \beta^*$ iff $\alpha \le \alpha^*$

Classification Error

$R(\hat{c}|\mathcal{Z}) = \sum_i I_{(\hat{c}(x_i)\neq y_i)}$
$R(\hat{c}) = \sum_{y\le k} p(y)E_{x|y}[I_{(\hat{c}(x_i)\neq y_i)}|Y = y]$ + terms from $\mathcal{D}$

Gradient Descent and Newton

Gradient descent:
$a(k+1) = a(k) - \eta_k \nabla J(a(k))$ Then using taylor second order expansion on $J(k)$ we have:
$J(a(k+1)) \approx J(a(k)) - \eta_k \nabla J^T J + \frac{1}{2}\eta_k^2 \nabla J^T H \nabla J$
Then $\eta^{opt} = \frac{\|\nabla J\|^2}{\nabla J^T H \nabla J}$ or we can use it another way as **Newton's method**:
$a(k+1) = a(k) - H^{-1}\nabla J(a(k))$. But this is hard to compute because of **hessian high dimentionality**.

Newton Exercise:
higher order roots: $\frac{f(x_n)}{f'(x_n)} = \frac{1}{\frac{k}{\epsilon_n} + \frac{g'(x_n)}{g(x_n)}}$ therefore $\epsilon_{n+1} = \epsilon_n(1 - \frac{1}{k+\epsilon_n\frac{g'(x_n)}{g(x_n)}})$ Then use the Taylor $\frac{1}{k+x} = \frac{1}{k} - \frac{x}{k^2} + O(x^2)$. How to change?
$x_{n+1} = x_n - k\frac{f(x_n)}{f'(x_n)}$

Perceptron

goal: $\forall x_i : a^T\tilde{x}_i \ge 0$ and cost function is: $J(a) = \sum_{\tilde{x}\in\mathcal{X}^{mc}} -a^T\tilde{x}$. Update rule in single sample mode is $a \leftarrow a + \tilde{x}$ if $\tilde{x}$ is mis-classified. If $\hat{a}$ exists as a solution for perceptron it can be proved that it **always converges**. Idea: $\|a(k+1) - \alpha\hat{a}\|^2$ shrinks depending on using proper α. How many iterations to converge?
$k_0 = \frac{max_{i\in\mathcal{X}^{mc}}\|\tilde{x_i}\|^2\|\hat{a}\|^2}{(min_{i\in\mathcal{X}^{mc}}\hat{a}^T x_i)^2}$ So for data almost orthogonal to a we need more iterations to converge.

Dual problem

$max\ inf_w \mathcal{L}(w,\lambda,\alpha)$ when $\forall\alpha_i : \alpha_i \ge 0$ Strong duality holds only by Slater's condition: f and g are convex and h is affine we should have: $h_j(w) < 0\ \forall j$.

SVM

Hard Margin

Minimize $\frac{1}{2}w^Tw$ s.t. $\forall i z_i(w^Ty_i + x_0) \ge 1$ Then dual is:
$\theta(\alpha) = \sum_{i\le n} a_i - \sum_{i\le n}\sum_{j\le n}\alpha_i\alpha_j z_i z_j y_i^T y_j$ s.t. $\forall i : \alpha_i \ge 0$ and $\forall i : \sum_{i\le n}\alpha_i z_i = 0$. Then after solving that we have optimal w as:
$w^* = \sum_{i\le n}\alpha_i^* z_i y_i$ (number of effective y_is are limited due to slack property. and
$w_0^* = \frac{-1}{2}[min_{i:z_i=-1}\ w^{*T}y_i + min_{i:z_i=1}\ w^{*T}y_i]$.
Optimal margin: $w^Tw = \sum_{i\in SV}\alpha_i^*$
KKT Conditions: only then strong duality holds: $\alpha_i^* \ge 0$
$\alpha_i^*(z_i g^*(y_i) - 1) = 0, (z_i g^*(y_i) - 1) \ge 0$

Soft Margin

Minimize $\frac{1}{2}w^Tw + C\sum_{i\le n}\xi_i$ s.t.
$\forall i z_i(w^Ty_i + x_0) \ge 1 - \xi_i,\ \xi_i \ge 0$. Exactly like before but we have also $0 \le \alpha \le C$ and if C is large we want less data points violate the margin. KKT condition has extra term $\xi_i(C - \alpha_i) = 0$. Data points with $\xi_i = 0$ are correctly classified and are either on the margin or on the correct side of the margin. Points for which $0 < \xi_i \le 1$ lie inside the margin, but on the correct side of the decision boundary, and those data points for which $\xi_i > 1$ lie on the wrong side of the decision boundary and are misclassified.

Multi Class SVM

$\underset{w,\xi\ge 0}{min}\frac{1}{2}w^Tw + C\sum_i\xi_i$ and $\frac{1}{2}w^Tw = \sum_z\frac{1}{2}w_z^Tw_z$. such that:
$\forall y_i \in \mathcal{Y} : {w_{z_i}}^T y_i + w_{z_i,0} - \underset{z\neq z_i}{max}(w_z^T y_i + w_{z,0}) \ge 1 - \xi_i$

SSVM

$\underset{w,\xi\ge 0}{min}\frac{1}{2}w^Tw + C\sum_i\xi_i$ such that: $\forall y_i \in \mathcal{Y}$:
$w^T\Psi(z_i,y_i) - \underset{z\neq z_i}{max} w^T\Psi(z,y_i) \ge \Delta(z,z_i) - \xi_i \Leftrightarrow$
$w^T\Psi(z_i,y_i) - \underset{z\neq z_i}{max}(w^T\Psi(z,y_i) + \Delta(z_i,z)) \ge -\xi_i$

Bagging and Boosting

Bagging

For comparing two bagged classifiers, first compare then bag! meaning: compare risks in each bootstrap samples, at the end look at the median. **Because sometimes the variability in bootstrap samples is much higher than the variability of the classifiers!**

Boosting

In this approach data everything is deterministic and diversity doesn't come from randomness, but from the weights that we assign to each data point. start with $w_0 = \frac{1}{n}$ and in iteration number b do the following:

1. $\epsilon_b \leftarrow \frac{\sum_i w_b^i I_{c_b(x_i)\neq y_i}}{\sum_i w_b^i}$
2. $\alpha_b \leftarrow log\frac{1-\epsilon_b}{\epsilon_b}$
3. $\forall i : w_b^i \leftarrow w_b^i exp(\alpha_b I_{c_b(x_i)\neq y_i})$

The final classifier is $\hat{c}_B(x) = sign(\sum_b \alpha_b c_b(x))$. This approach is equivalent to minimizing the exponential loss $exp(-y\hat{F}_B(x))$

PAC Learning

Definition: Algorithm $\mathcal{A}$ can learn a concept c if for sample size $n \ge poly(\frac{1}{\epsilon}, \frac{1}{\delta}, dim(\mathcal{X}))$ we have: $\mathbb{P}_{Z\sim\mathcal{D}^n}(\mathcal{R}(\hat{c}) \le \epsilon) > 1 - \delta$. If $\mathcal{A}$ runs polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$ then c is efficiently PAC learnable. if labels are not deterministic:
$\mathbb{P}_{Z\sim\mathcal{D}^n}(\mathcal{R}(\hat{c}) - inf_{c\in\mathcal{C}R(c)} \le \epsilon) > 1 - \delta$.
If we have uniform convergence for $\hat{R}(\hat{c}_n^*)$ to $R(c_n^*)$ and point-wise convergence for $\hat{R}(c^*)$ where c^* is the best possible classifier, then we can show: $\mathbb{P}\{\mathcal{R}(\hat{c}_n^*) - \mathcal{R}(c^*) > \epsilon\} \le \mathbb{P}\{sup_{c\in\mathcal{C}}|\hat{\mathcal{R}}(c) - \mathcal{R}(c)| > \epsilon/2\}$
If we train a classifier $\hat{c}_n^*(x)$ on dataset $\mathcal{Z}$ the **generalization error** is $\mathbb{P}(\hat{c}_n^*(X) \neq Y|\mathcal{Z})$
Hoeffding's Inequality: (Markov) if X is a non-negative r.v. : $\mathbb{P}\{X \ge \epsilon\} \le \frac{\mathbb{E}[X]}{\epsilon}$
If $\mathbb{E}[X] = 0$ and $a \le X \le b$ then:
$\mathbb{E}[e^{sX}] \le exp(s^2(b-a)^2/8)$. How to prove? notice $e^{sX} \le e^{sb}\frac{x-a}{b-a} + e^{sa}\frac{b-x}{b-a}$ then let $p := \frac{-a}{b-a}$, $u := s(b-a)$ and notice the right hand side is equal to
$e^{\Phi(u)}, \Phi(u) = -pu + log(1 - p + pe^u)$. Then by using Taylor expansion and approx we reach the result. Final result:
$\mathbb{P}\{s_n - \mathbb{E}[S_n] \ge t\} \le exp(-\frac{2t^2}{\sum_i(b_i-a_i)^2})$ and
$\mathbb{P}\{s_n - \mathbb{E}[S_n] \le -t\} \le exp(-\frac{2t^2}{\sum_i(b_i-a_i)^2})$.

Useful Lemma

$\mathbb{P}(a + b > \epsilon) \le \mathbb{P}(a > \epsilon/2) + \mathbb{P}(b > \epsilon/2)$
$\mathbb{P}(\bigcup_i A_i) = \sum_i A_i - \sum_{1\le i_1<i_2\le n}\mathbb{P}(A_{i_1}\cap A_{i_2}) + \dots + (-1)^{n-1}\mathbb{P}(A_1,\dots,A_n)$

Error Bounds for Finite Classifier Sets

$\mathbb{P}\{sup_{c\in\mathcal{C}}|\hat{\mathcal{R}}_n(c) - \mathcal{R}(c)| > \epsilon\} \le 2|\mathcal{C}|exp(-n\epsilon^2)$
Then we can say:
$\mathcal{R}(c) \le \hat{\mathcal{R}}_n(c) + \sqrt{\frac{ln|\mathcal{C}| - ln(\sigma/2)}{2n}}$

Error Bounds for Hyperplanes in $\mathbb{R}^d$

$\mathbb{P}\{\mathcal{R}(\hat{c}) - \mathcal{R}(c^*) > \epsilon\} \le (2\binom{n}{d} + 1)e^{2d\epsilon}e^{-n\epsilon^2/2}$

Beta and Dirichlet Distribution

$Beta(x|a,b) = \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}$, $x \in [0,1], a, b > 0$ and $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$,
$\Gamma(a) = \int_0^\infty e^{-x}x^{a-1}dx$
$Dir(x|\alpha) = \frac{1}{B(\alpha)}\prod_k x_k^{a_k-1}$ where $x = [x_1, x_2, \dots, x_n]$, $x_k \in [0,1]$, $\alpha_k > 0$. $B(\alpha)$ is a generalization of beta function meaning
$B(\alpha) = \frac{\prod_k\Gamma(a_k)}{\Gamma(\sum_k a_k)}$

Clustering

Purity Exercise:

$$\begin{aligned}
I(\mathcal{U},\mathcal{V}) &= \sum_{i=1}^R\sum_{j=1}^C p_{UV}(i,j)\log_2\frac{p_{UV}(i,j)}{p_U(i)p_V(j)} \\
&= \sum_{i=1}^R\sum_{j=1}^C p_{UV}(i,j)\frac{p_V(j)}{p_V(j)}\log_2\frac{p_{UV}(i,j)}{p_U(i)p_V(j)} \\
&= \sum_{i=1}^R\sum_{j=1}^C p_V(j)\frac{p_{UV}(i,j)}{p_V(j)}\left(\log_2\frac{p_{UV}(i,j)}{p_V(j)} - \log_2 p_U(i)\right) \\
&= \sum_{j=1}^C p_V(j)\sum_{i=1}^R\frac{p_{UV}(i,j)}{p_V(j)}\left(\log_2\frac{p_{UV}(i,j)}{p_V(j)} - \log_2 p_U(i)\right) \\
&= \sum_{j=1}^C p_V(j)\left(-H(\mathcal{U}\mid\mathcal{V}) - \sum_{i=1}^R\frac{p_{UV}(i,j)}{p_V(j)}\log_2 p_U(i)\right) \\
&= -H(\mathcal{U}\mid\mathcal{V}) - \sum_{j=1}^C\sum_{i=1}^R p_{UV}(i,j)\log_2 p_U(i) \\
&= -H(\mathcal{U}\mid\mathcal{V}) - \sum_{i=1}^R p_U(i)\log_2 p_U(i) \\
&= H(\mathcal{U}) - H(\mathcal{U}\mid\mathcal{V})
\end{aligned}$$

Jensen's Inequality:
$-\sum_i\sum_j p_{UV}(i,j)log\frac{p_U(i)p_V(j)}{p_{UV}(i,j)} \ge -log\sum_i\sum_j p_{UV}(i,j)\frac{p_U(i)p_V(j)}{p_{UV}(i,j)}$

De Finetti's Theorem
if $(X_1,\dots,X_n)$ n infinitely exchangeable sequence of random variables. Then $\forall n$:
$P(X_1,\dots,X_n) = \int(\prod_i p(x_i|G))dP(G)$

Gibbs Sampling
$p(z_i = k|z_{-i}, x, \mu, \alpha) \propto p(z_i = k|z_{-i}, \alpha)p(x_i|x_{-i}, z_i, z_{-i}, \mu)$
$p(z_i = k|z_{-i}$ = prior $\times$ likelihood:

$$\begin{cases} \frac{N_{k,-i}}{\alpha+N-1}\frac{p(x_i,x_{-i,k}|\mu)}{p(x_{-i,k}|\mu)} & \text{for existing k} \\ \frac{\alpha}{\alpha+N-1}p(x_i|\mu) & \text{for new k} \end{cases}$$

for $i = 1$ to N in random order **do**
  Remove x_i's sufficient statistics from old cluster z_i;
  for $k = 1$ to K **do**
    Compute $p_k(x_i) = p_k(x_i|\boldsymbol{x}_{-i,k})$;
    Set $N_{k,-i} = |\boldsymbol{x}_{-i,k}|$;
    Compute $p(z_i = k|\boldsymbol{z}_{-i}, \boldsymbol{x}) = \frac{N_{k,-i}}{\alpha+N-1}$;
  end for
  Compute $p_\star(x_i) = p(x_i|\boldsymbol{\mu})$;
  Compute $p(z_i = \star|\boldsymbol{z}_{-i}, \boldsymbol{x})$;
  Normalize $p(z_i|\cdot)$;
  Sample $z_i \sim p(z_i|\cdot)$;
  Add x_i's sufficient statistics to new cluster z_i;
  If any cluster is empty, remove it and decrease K;
end for

Latent Dirichlet Allocation
Distribution of topics in document d: $\theta_d \sim Dir(\alpha)$
What topic the word w belongs to in document d: $z_{d,w} \sim$ Categorical(θ_d)
Distribution of words in topic k:$\Phi_k \sim Dir(\beta)$
What is the word w in document d:
$w_{d,w} \sim$ Categorical$(\Phi_{z_{d,w}})$