Bregman $d_f(x,y) = f(x) - f(y) - (x-y)^T \nabla f(y)$. If $f(x) = \|x\|_2^2$: $d_f = \|.\|_2$. If $f(x) = \sum_i x_i \log x_i - x_i$, $KL(x,y) = \sum x_i \log \frac{x_i}{y_i} - (x_i - y_i)$.

Fitting Gaussians $x \in R^D$. $N(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$. If $\Sigma_1 = \Sigma_2$: $(m_2 - m_1)^T \Sigma x + c = 0$ is decn surface.

Opt Take primal $\min f_0(x) : \{f_i(x) \leq 0\}, \{h_i(x) = 0\}$; Get Lagrangian: $L(x,l,m)$; get $g(x) = \inf_x L(x,l,m)$; Solve $\max_{l,m} g(l,m)$; derive x^* from l^*, m^*. **KKT:** Primal feasibility: $f(x^*) \leq 0, h(x^*) = 0$. Dual feasiblity: $l^* \geq 0$. Complimentary slackness: $\forall j : l_j^* f_j(x^*) = 0$. Optimality: $x^* = argmin_x L(x, l^*, m^*)$: set $\nabla_{x^*} L(x, l^*, m^*) = 0$.

SVM $c(x) = sgn(\frac{w^T x + w_0}{\|w\|})$. $\max_{w,w_0} [\frac{\min_n [y(x_n)c(x_n)]}{\|w\|}]$. Scale w, w_0 so that $\min_n [y(x_n)c(x_n)] = 1$; thence get $\equiv$ problem $\min_{w,w_0} \frac{\|w\|^2}{2}$: $y(x_n)c(x_n) \geq 1$. Prediction: sgn(y(x)). Get Lagrangian $L(w, w_0, a) = \frac{\|w\|^2}{2} + \sum a_n [1 - (w^T \phi(x_n) + w_0) c(x_n)]$; $a_n \geq 0$. Dual: $\max_a g(a) = \max \sum a_n - 2^{-1} \sum_n \sum_m a_n a_m c(x_n) c(x_m) k(x_n, x_m)$: $a_n \geq 0; \sum a_n c(x_n) = 0$. Predictor: $y(x) = \sum_n a_n c(x_n) k(x_n, x) + w_0$. $w_0 = \frac{\sum_m [c(x_m) y(x_m) - \sum_n a_n k(x_n, x_m)]}{N}$.

Soft SVM $\min C \sum_{n=1}^N \xi_n + \frac{\|w\|^2}{2}$: $\xi_n \geq 0$; $y(x_n)c(x_n) + \xi_n \geq 1$. Same dual, but constraints: $0 \leq a_n \leq C$: as $\mu_n \geq 0; \sum a_n c(x_n) = 0$. Complimentary slackness: $a_n(1 - c(x_n) y(x_n) - \xi_n) = 0, \mu_n \xi_n = 0$.

Logistic k-class problem. Model: $\forall i \in [1 : k-1] : \log \frac{Pr(C_i|x)}{Pr(C_k|x)} = w_{i0} + w_i^T x$. Get: $Pr(C_i|x) = \frac{e^{w_{i0} + w_i^T x}}{1 + \sum e^{w_{j0} + w_j^T x}}, Pr(C_k|x) = \frac{1}{1 + \sum e^{w_{i0} + w_i^T x}}$! 2 class: $\min E(w) = \sum l_i \log(\frac{1}{1 + e^{w^T x_i}}) + \sum (1 - l_i) \log(1 - \frac{1}{1 + e^{w^T x_i}})$.

LDA Before projection: Take $S_T = \sum_x (x-m)(x-m)^T$; $S_W = \sum_{i=1}^k \sum_{x \in C_i} (x - m_i)(x - m_i)^T$; $S_B = \sum_{i=1}^k n_i (m_i - m)(m_i - m)^T$. So, $S_T = S_W + S_B$.

After projection scatters: $S'_W = W^T S_W W, S'_B = W^T S_B W$. Find $\max_W \frac{|W^T S_B W|}{|W^T S_W W|}$ or maybe $\max_W tr((W^T S_W W)^{-1} (W^T S_B W))$. same as ev problem $S_W^{-1} S_B x = \lambda x$.

Perceptron Update: $w_{t+1} = w_t + y_t x_t$; Min margin: $y_t (w^{*T} x_t) \geq \gamma$; $\|x_i\|^2 \leq R^2$, $w_0 = 0$. Convergence: $w^{*T} w_t \geq t\gamma$, $\|w_t\|_2^2 \leq tR^2$.

k means $S' = (S'_i) = argmin_S \sum_{i=1}^k \sum_{x_j \in S_i} d(x_j, \mu_i)$. If d is any Bregman div, k means minimizes this at each iteration: Alg finds better clustering, Mean is best cluster representative.

Least squares $x_i \in R^d$. $w_0 = \bar{y} - \sum_{i=1}^d w_i \bar{x_j}$.