## Polynomial Regression

- feature extraction $-\ \phi(x) = [x_i^0, .., x_i^M]^T \in \mathbb{R}^{1x(M+1)}$
- mapper $-\ y(x_i, w) = \sum_{m=0}^{M} w_m x_i^m = Xw$
- objective
  - $J(w) = \frac{1}{2}\sum_{i=1}^{N}(t_i - y_i)^2 = \frac{1}{2}\sum_{i=1}^{N}(t_i - Xw)^T(t_i - Xw)$
  - $\frac{\partial J}{\partial w} = 0 \rightarrow w = (X^T X)^{-1} X^T t$

## Regularization

- Ridge $-\ R_w^{L2} = \lambda \sum_{m=0}^{M} w_m^2$
- Lasso $-\ R_w^{L1} = \lambda \sum_{m=0}^{M} |w_m|$
- Elastic Net $-\ R_w^{L12} = \beta\lambda \sum_{m=0}^{M} |w_m| + (1-\beta)\lambda \sum_{m=0}^{M} w_m^2$
- regularized objective with ridge
  - $||t - Xw||_2^2 + \lambda ||w||_2^2$
- regularized weight
  - $\frac{\partial J}{\partial w} = 0 \rightarrow w = (X^T X + \lambda I)^{-1} X^T t$

## Least Squares Estimators

- MLE (frequentist)
  - $w_{MLE} = \arg_w \max P(t|w)$
- MAP (Bayesian)
  - $w_{MAP} = \arg_w \max P(t|w)P(w)$

## Bayesian Interpretation of Objective

- $\arg_w \min J(w) = \arg_w \max -J(w) = \arg_w \max \exp(-J(w))$

## Performance Measures

* accuracy = correct/total
* error = wrong/total
* precision (P) = TP/(FP+TP)
* recall (R) = TP/(TP+FN)
* specificity = TN/(TN+FP)
* fallout = FP/(FP+TN)
* F1 = 2*P*R/(P+R)
* MSE $= \frac{1}{n}\sum_{i=1}^{N}(t_i - y_i)^2$
* MAE = $= \frac{1}{n}\sum_{i=1}^{N}|t_i - y_i|$

## Gaussian Mixture Model

- $P(x|\theta) = \sum_{k=1}^{K} \pi_k P(x|\theta_k) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)$
- observed data likelihood $-\ L^0 = \prod_{i=1}^{N}\sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \Sigma_k)$
- complete data likelihood $-\ L^C = \prod_{i=1}^{N} \pi_{z_i} N(x_i|\mu_{z_i}, \Sigma_{z_i})$
- latent variable (gaussian component from which $x_n$ was drawn from) $z_n = \{1,...,K\}$
- membership of $x_i$ in component
  - $C_{ik} = P(z_i|x_i, \theta) = \frac{P(x_i|\mu_{z_i}, \Sigma_{z_i})\pi_{z_i}}{\sum_{k=1}^{K} P(x_i|\mu_k, \Sigma_k)\pi_k}$
- optimization $-\ Q(\theta, \theta^t) = \sum_{z_i=1}^{K} \ln(L^c) P(z_i|x_i, \theta^t)$
- E-step $-\ P(z_i|x_i, \theta^t)$
- M-step $-\ Q(\theta, \theta^t) = \sum_{z_i=1}^{K} \ln(L^c) P(z_i|x_i, \theta^t)$
- $\mu_k = \frac{\sum_{i=1}^{N} x_i C_{ik}}{\sum_{i=1}^{N} C_{ik}}$, $\sigma_k = \frac{\sum_{i=1}^{N} C_{ik}||x_i - \mu_k||_2^2}{d\sum_{i=1}^{N} C_{ik}}$, $\pi_k = \frac{\sum_{i=1}^{N} C_{ik}}{N}$
- $Q_\pi(\theta, \theta^t) = Q(\theta, \theta^t) + \lambda(1 - \sum_{k=1}^{K} \pi_k)$

## K-means

- $J(\theta, U) = \sum_{i=1}^{N}\sum_{k=1}^{K} u_{ik} d^2(x_i, \theta_k)$ where $u_{ik} \in \{0,1\}$
- $\theta_k = \frac{\sum_{x \in C_k} x}{N_k}$
- optimization $-\ \arg_{\theta,U} \min J(\theta, U)$

## Bayes Theorem

$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

## Bayes Law of Total Probability

$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{\sum_{i=1}^{N} P(B|A_i)P(A_i)}$

## Cluster Validity

- Silhouette $-\ \frac{1}{N}\frac{\sum_{i=1}^{N} b_i - a_i}{\max(b_i, a_i)}$, a = avg dist within cluster, b = avg dist to other cluster
- Rand $-\ \frac{a+b}{a+b+c+d}$, a+b = # of agreements, c+d = # of disagreements

## Naive Bayes Classifier

- $P(C_k) = \frac{N_k}{N}$

## Distributions

- Laplacian $-\ f(x; \mu, b) = \frac{1}{2b}\exp(-\frac{|x-\mu|}{b})$
- Gamma $-\ f(x; \alpha, \beta) = \frac{x^{\alpha-1}\exp(-\beta x)\beta^\alpha}{\Gamma(\alpha)}$ where $\Gamma(\alpha) = (\alpha - 1)!$
- Bernoulli $-\ P(x; p) = p^x(1-p)^{1-x}$
- Beta $-\ f(x; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$
- Poisson $-\ P(x) = \frac{\lambda^x \exp(-\lambda)}{x!}$
- Gaussian $-\ f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$
- Multivariate Gaussian $-$ $f(x; \mu, \sigma) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2})$

## Distance Metrics

- Euclidean $-\ ||x_1 - x_2||_2$
- Mahalanobis $-\ \sqrt{(x_1 - x_2)^T\Sigma^{-1}(x_1 - x_2)}$
- Cosine $-\ 1 - \frac{(x_1^T x_2)}{||x_1||_2||x_2||_2}$

## Linear Algebra

* $(AB)^T = B^T A^T$
* $(AB)^{-1} = B^{-1}A^{-1}$
* $\det(A) = \det\begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$

## Calculus

- Chain rule $-\ (f \circ g)'(x) = f'(g(x))g'(x)$

**Basis Functions**

- Radial $-\ \phi_j(x) = \exp(-\frac{(x-\mu_j)^2}{2s^2})$
- Sigmoidal $-\ \phi_j(x) = \sigma(\frac{x-\mu_j}{s})$ where

$$\sigma(a) = \frac{1}{1+\exp(-a)}$$

**Bias variance Tradeoff**

$E[(t-y)^2]$ = variance + bias$^2$ + error