$E[(Y - \hat{f}(x))^2 \mid X = x] = [f(x) - \hat{f}(x)]^2 + Var(\epsilon)$

MSE = $1/n \sum (y_i - \hat{f}(x))^2 = E[(Y - \hat{f}(x))^2]$

$E[(y_0 - \hat{f}(x_0))^2] = Var(\hat{f}(x_0)) + (f(x_0) + E(\hat{f}(x_0)))^2 + Var(\epsilon)$

Classification Error (Loss) = $1/n \sum I[y_i \neq \widehat{C(x_i)}]$

RSS = $\sum(y_i - \widehat{\square}_0 - \widehat{\square}_1 x_i)^2$ -- Matrix form $(Y - x\beta)^T (Y - x\beta) = \epsilon^T \epsilon$

TSS = $\sum(y_i - \bar{y})^2$; $R^2 = \frac{TSS - RSS}{TSS}$;RSE = $\sqrt{\frac{RSS}{n-p}}$

$\beta_0 = \bar{y} - \beta_1 \bar{x}$; $\beta_1 = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2} = \frac{\Sigma x_i (y_i - \bar{y})}{\Sigma x_i (x_i - \bar{x})}$

$SE(\widehat{\beta_0})^2 = \sigma^2 [\frac{1}{n} + \frac{\overline{x^2}}{\Sigma(x_i - \bar{x})^2}]$, $SE(\widehat{\beta_1})^2 = \sigma^2 [\frac{1}{n} + \frac{\sigma^2}{\Sigma(x_i - \bar{x})^2}]$, where $\sigma^2 = Var(\epsilon)$;

95% Confidence Interval for $\beta_1 = [\widehat{\square}_1 - 2 * SE(\widehat{\square}_1), \widehat{\square}_1 + 2 * SE(\widehat{\square}_1)]$

$\widehat{\sigma^2} = \frac{RSS}{n-2}$; $\square = (x^T x)^{-1} x^T y \leftarrow$ OLS Estimate

$Var(\square) = (x^T x)^{-1} \sigma^2 \leftarrow$ Matrix Form | Non-Matrix Form $\rightarrow SE(\widehat{\beta_1})^2 = \frac{\sigma^2}{\Sigma (x_i - \bar{x})^2}$

$\beta_{P\times1} = \text{argmin}_\beta (y_i - x\beta)^T (y_i - x\beta)$; T-statistic $(\beta_1) = \frac{\widehat{\beta_1}}{SE(\widehat{\beta_1})}$

Correlation Coefficient: $r = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma (x_i - \bar{x})^2} \sqrt{\Sigma (y_i - \bar{y})^2}}$; F-statistic = $\frac{(TSS - RSS)/p}{RSS/(n-p-1)}$

Logistic Regression $p(x) = \frac{e^{(\beta_0 + \beta_1 X)}}{1 + e^{(\beta_0 + \beta_1 X)}}$

If $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ and $\epsilon_i \sim N(0, \sigma^2)$

$$\ell = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi\sigma^2}} e^{(-\frac{(\epsilon_i - 0)^2}{2\sigma^2})} = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi\sigma^2}} e^{(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2})}$$

Maximum Likelihood Estimate = Min RSS

Bayes Theorem: Pr(Y=k|X=x) = (Pr(X=x|Y=k) • Pr(Y=k)) / Pr(X=x)

Bayes Classifier: $p_k(x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$ and $log(p_k(x)) = log(\pi_k) + log(f_k(x)) + const$

$\widehat{\delta\square}(x) = x \bullet \frac{\widehat{\mu\square}}{\widehat{\sigma^2}} - \frac{\widehat{\mu\square}}{\widehat{2\sigma^2}} + log(\widehat{\pi\square})$ <- LDA Single-variate Discriminant

$\delta_k = x^T \Sigma^{-1} \mu_k - \mu_k^T \Sigma^{-1} \mu_k + log\, \pi_k$ <- LDA Multivariate Discriminant

$\delta_k = -\frac{1}{2} x^T \Sigma_k^{-1} x + x^T \Sigma_k^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k - \frac{1}{2} log|\Sigma_k| + log\, \pi_k$ <- QDA

$$Pr(y = k | x = x) = \frac{e^{\delta_k(x)}}{\sum_{\ell}^{K} e^{\delta_\ell(x)}}$$

Maximum Log Likelihood QDA = max $\sum log(\frac{1}{\sqrt{2\pi\sigma^2}} e^{(-\frac{(x_i - \mu_{yi})^2}{2\sigma^2})})$ = max $\sum_{i=1}^{N} (-\, log\, \sigma - \frac{1}{2\sigma^2}(x_i - \mu_{y_i})^2 + consts)$

X = [a b, c d]; det(X) = ad - bc; X^-1 = (1/detX)[d -b, -c a]
↑Flexibility → ↑Var + ↓Bias

Confusion Matrix	**True**	
Pred	Correct	Type 1(False +)
	Type 2(False -)	Correct

FPR: FP / (TN + FP); (type 1)
TPR: TP / (TP + FN)

C(x) : j, if $p_j(x) = max\{p_1(x),.....,p_k(x)\}$, where $p_k(x) = Pr(Y=k \mid X=x)$

Multivariate Gaussian PDF:

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

Logistic regression model: $log(\frac{p(x)}{1-p(x)}) = \beta_0 + \beta_1 X$

Log reg. PDF: $p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1+e^{\beta_0 + \beta_1 x}}$

Log Likelihood Log reg.: $log\, \ell(x) = log(\Pi_{i=1}^{N} (p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}))$

Var(X+Y) = Var(X)+Var(Y) + 2Cov(X,Y); Var(aX) = a^2Var(X); Cov(aX,bY) = abCov(X,Y);
TSS = $\sqrt{RSE/(n-2)}/(1-R^2)$; LOOCV: CV = 1/n ∑ MSE
Speeding up LOOCV: CV = 1/n ∑ $((y_i - \widehat{y_i})/(1 - h_i))^2$

$h_i = 1/n + ((x_i - \bar{x})^2 / \Sigma (x_j - \bar{x})^2)$

K-fold cv: $\sum n_k/n$ MSE

classification cv: $\sum n_k/n$ Err

Err = $\sum I(y_i \neq \widehat{y_i})/n_k$

$$C_p = \frac{1}{n}\left(\text{RSS} + 2d\hat{\sigma}^2\right), \quad \text{AIC} = -2\log L + 2\cdot d$$

L is likelihood function

$$\text{BIC} = \frac{1}{n}\left(\text{RSS} + \log(n)d\hat{\sigma}^2\right).$$

Maximum Log Likelihood = max $\sum log(\frac{1}{\sqrt{2\pi\sigma^2}}e^{(-\frac{(x_i-\mu_{yi})^2}{2\sigma^2})})$

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}.$$

- **Ridge regression** (parameter λ), ℓ_2 penalty

$$\min_\beta \text{RSS}(\beta) + \lambda \sum_j \beta_j^2 =$$

$$\min_\beta \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij}\right)^2 + \lambda \sum_j \beta_j^2$$

Constrained Ridge

$$\hat{\beta}^{\text{ridge}} = \arg\min_\beta \|Y - X^T\beta\|^2 + \lambda\|\beta\|^2,$$

Another way to formulate the ridge regression is

$$\hat{\beta}^{\text{ridge}} = \arg\min_\beta \|Y - X^T\beta\|^2$$

$$\text{subject to } \|\beta\|^2 \le t,$$

For ridge regression: $\hat{\beta} = (X^TX + \lambda I)^{-1}X^TY$

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^n (x_{ij} - \overline{x}_j)^2}}$$

<= Standardized predictors

$$\min_\beta \text{RSS}(\beta) + \lambda \sum_j |\beta_j| =$$

<= Lasso

$$\min_\beta \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij}\right)^2 + \lambda \sum_j |\beta_j|$$

Constrained Lasso => $$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \le s$$

Polynomial => $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_p X^P$

Step function => $Y = \beta_0 + \beta_1 I(X < c_1) + \beta_2 I(c_1 < X < c_2) + \dots + \beta_p I(c_p < X)$

Cubic Spline => $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^2 + \beta_4 (X_i - \xi_1)_+ + \dots + \beta_{k+3} (X_i - \xi_k)_+$

Smoothing Spline => $$\underset{g \in \mathcal{S}}{\text{minimize}} \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

Gini Index => $G = \sum_{k=1}^{K} \hat{P}_{mk}(1 - \hat{P}_{mk})$

Cross-Entropy => $D = -\sum_{k=1}^{K} \hat{P}_{mk} log_2(\hat{P}_{mk})$

If $Var(\alpha X + (1 - \alpha)Y)$then minimizing this is $\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$

SVM (Hard-Margin)

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,M}{\text{maximize}} \; M$$
$$\text{subject to} \; \sum_{j=1}^{p} \beta_j^2 = 1,$$
$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \ge M \quad \forall \, i = 1, \ldots, n.$$

SVM (Soft-Margin)

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n,M}{\text{maximize}} \; M$$
$$\text{subject to} \; \sum_{j=1}^{p} \beta_j^2 = 1,$$
$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \ge M(1 - \epsilon_i),$$
$$\epsilon_i \ge 0, \quad \sum_{i=1}^{n} \epsilon_i \le C, \qquad \epsilon_1, \ldots, \epsilon_n \text{ are } \textit{slack variables}$$

Linear support vector classifier => $$f(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p = \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\alpha}_i \langle x, x_i \rangle$$

Kernel version ^^ => $$f(x) = \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\alpha}_i k(x, x_i)$$

Polynomial Kernel => $$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d$$

Radial Kernel => $$K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2).$$

SVM Hinge Loss => $$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{minimize}} \left\{ \sum_{i=1}^{n} \max\left[0, 1 - y_i f(x_i)\right] + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

PCA and Eigen: $maximize\ \phi$ with $\phi_1{}^T \Sigma \phi_1\ subject\ to\ \phi_1{}^T \phi_1 = 1$which is equivalent to $\phi_1{}^T \Sigma \phi_1 - \lambda(\phi_1{}^T \phi_1 - 1)$ and then taking gradient and setting to zero ends up being λ

PCA: $Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + ... + \phi_{p1} X_p$ for single component

$$\underset{\phi_{11},\dots,\phi_{p1}}{\text{maximize}} \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \text{ subject to } \sum_{j=1}^{p} \phi_{j1}^2 = 1.$$

$Z_{nx1} \Phi_1{}^T = x^*{}_{nx2} = denoised\ data$

$\hat{\Sigma} = \frac{X^T X}{n}$

$\sum_{j=1}^{p} Var(x_j) = \sum_{m=1}^{M} Var(Z_m)$

K-means:

$min\ C_1 ... C_k \sum_{k=1}^{K} WCV(C_k)\ where\ WCV\ (C_k) = \frac{1}{|C_k|} \sum_{i,i' \epsilon C_k} \sum_{j=1} (x_{ij} - x_{i'j})^2 = \frac{1}{|C_k|} \sum_{i \epsilon C_k} \sum_{j=1} (x_{ij} - \bar{x}_{kj})^2$

As K increases, cost function decreases