STA9890 - Test 2 - Formula Sheet

Ordinary Least Squares:

- $\hat{\beta} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$

Penalties:

- Best Subsets: $\|\boldsymbol{\beta}\|_0$. Number of non-zero elements (non-convex), induces sparsity in $\boldsymbol{\beta}$
- Lasso: $\|\boldsymbol{\beta}\|_1$. Tightest convex relaxation of best subsets
- Ridge: $\frac{1}{2}\|\boldsymbol{\beta}\|_2^2$. Very nice to work with (differentiable). $\frac{\partial}{\partial \boldsymbol{\beta}}\|\boldsymbol{\beta}\|_2^2 = 2\boldsymbol{\beta}$.
- Elastic Net: α-weighted combination of ridge and lasso $\alpha\|\boldsymbol{\beta}\|_1 + \frac{(1-\alpha)}{2}\|\boldsymbol{\beta}\|_2^2$

Classification:

- (True/False) (Positive/Negative) = (Correct/Incorrect) Prediction
- Generative: $p(X|Y) \implies p(Y|X)$ via prior and Bayes' Rule. Discriminative: model $p(Y|X)$ directly.
- Bayes' Rule:

$$p(A|B) = \frac{p(B|A) * p(A)}{P(B)} = \frac{p(B|A) * p(A)}{P(B|A) * P(A) + P(B|A^c) + P(A^c)}$$

Non-Linearity:

- Feature expansion and engineering: fit linear models to non-linear parts
- Splines: piecewise polynomial models with additional smoothness constraints
- Kernel methods: feature expansion made 'easy'. Replace inner product with a 'kernel function'.

Ensembles:

- Stacking: Linear combination of base learners. Typically non-negative and sum-to-one constrained
- Boosting: building an ensemble by averaging bootstrapped based learners
- Boosting: building an ensemble by adding new ensemble members to correct past mistakes. Fit slowly for 'gradient descent' on functions

Distributions:

- Bernoulli Distribution: $X \sim \text{Bernoulli}(p) \in \{0, 1\} \implies \mathbb{P}(X = x) = p^x(1-p)^{1-x}$
- Binomial Distribution: $X \sim \text{Binomial}(n, p) \in \{0, \dots, n\} \implies \mathbb{P}(X = x) = \binom{n}{x}p^x(1-p)^{n-x}$
- Poisson Distribution: $X \sim \text{Poisson}(\lambda) \in \{0, 1, \dots\} \implies \mathbb{P}(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$
- Standard normal distribution. $Z \sim \mathcal{N}(0, 1)$. Mean Zero + Variance 1
- Standard normal PDF - $\phi(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$. Standard normal CDF $\Phi(z) = \int_{-\infty}^{z} \phi(x)\,\mathrm{d}x$ - no closed form.
- General normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ - generated by scale+shift of standard normal $X \stackrel{d}{=} \mu + \sigma Z$.
- Normal PDF via standardization (z-score): $f_X(x) = \phi(\frac{x-\mu}{\sigma}) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2\sigma^2}$. CDF: $\Phi(\frac{x-\mu}{\sigma})$.
- Multivariate normal parameterized by mean vector and (co)variance matrix: $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Standard multi-normal: $\boldsymbol{Z} \sim \mathcal{N}_n(\boldsymbol{0}_n, \boldsymbol{I}_n)$. PDF $f_{\boldsymbol{Z}}(\boldsymbol{z}) = (2\pi)^{-n/2}e^{-\|\boldsymbol{z}\|^2/2}$.
- General multi-normal $\boldsymbol{X} \stackrel{d}{=} \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}$ where $\boldsymbol{\Sigma}^{1/2}$ is a matrix square root (Cholesky or symmetric).
- Bivariate normal PDF

$$f_{(X,Y)}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2\left[1-\rho^2\right]}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right)$$

- Multivariate normal: any linear combination (weighted sum) of X_i is normal.
- If $\mathbb{C}[X_i, X_j] = 0$, then $X_i \perp\!\!\!\perp X_j$ (for multi-normal, uncorrelated implies independent)
- If $\boldsymbol{Z}$ is a standard normal n-vector, $\|\boldsymbol{Z}\|^2 = \sum_{i=1}^{n} Z_i^2$ has a χ^2 distribution with n degrees of freedom