1 Definitions

1.1 Classifiers, Risk etc.

Definition 1. Bayes Risk is $R^\star = \inf_f R(f) = \inf_f \mathbb{E}[l(f,X,Y)] \overbrace{=}^{0-1 loss} \inf_f \mathbb{E}\left[\mathbb{1}_{f(X)\neq Y}\right] = \mathbb{P}\{f(X) \neq Y\}$.

Definition 2. Optimal Bayes Classifier is $f^\star(x) = 1$ if $\eta(x) \geqslant \frac{1}{2}$ and 0 otherwise; and equivalently $f^\star(x) = 1$ if $\frac{\eta(x)}{1-\eta(x)} \geqslant 1$ and 0 otherwise; and equivalently $f^\star(x) = 1$ if $\frac{p(x|Y=1)\,p(Y=1)}{p(x|Y=0)\,p(Y=0)} \geqslant 1$ and 0 otherwise.

Definition 3. Log Likelihood Ratio: $\Lambda(x) = \log\left(\frac{p_1(x)}{p_0(x)}\right)$, where $p_j(x) = p(x|Y=j)$.

Definition 4. Bayes Cost: $C = \sum_{i,j=0}^{1} c_{i,j}\pi_j \mathbb{P}\{\text{ decide } H_i|H_j\} = \sum_{i,j=0}^{1} c_{i,j}\pi_j \int_{R_i} p_j(x)\,dx$, where $\pi_j = \mathbb{P}\{H_j\}$ and $R_j = \{x: \text{ decide } H_j\}$.

Definition 5. MLE Risk: $R_{MLE}(q,p_\theta) = \mathbb{E}[-\log p(x|\theta)]$, Excess Risk: $R_{MLE}(q,p_\theta) - R_{MLE}(q,q) = D(q\|p_\theta) \geqslant 0$.

1.2 Estimators

Definition 6. Empirical means and covariances: $\hat{\mu}_j = \frac{1}{\#\{y_i=j\}} \sum_{i:y_i=j} x_i$ and $\hat{\Sigma} = \frac{1}{n}\left(\sum_j \sum_{i:y_i=j} (x_i - \hat{\mu}_j)(x_i-\hat{\mu}_j)^T\right)$.

Definition 7. Gaussian GLM: $p\left(y|x^T w\right) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-1}{2}\left(y - x^T w\right)^2\right)$, and $\hat{w} = \left(X^T X\right)^{-1} X^T y$.

Definition 8. Binomial GLM: $p\left(y|x^T w\right) = \exp\left(y \log\left(\frac{1}{1+e^{-x^T w}}\right) + (1-y)\log\left(\frac{1}{1+e^{x^T w}}\right)\right)$.

Definition 9. Multinomial GLM: $p\left(y|x^T w\right) = \frac{\exp\left(x^T w_l\right)}{\sum_{j=1}^{k} \exp\left(x^T w_j\right)}$.

Definition 10. Max a Posterior: $\theta_{MAP} = \max_\theta p(\theta|y) \propto p(y|\theta)\,p(\theta)$ to minimize loss $L\left(\theta,\hat{\theta}\right) = \mathbb{1}_{\{\|\hat{\theta}-\theta\|>\varepsilon\}}$.

Definition 11. Bayesian minimum MSE estimator: $\hat{\theta} = \mathbb{E}[\theta|y]$ to minimize loss $L\left(\theta,\hat{\theta}\right) = \|\theta - \hat{\theta}\|^2$.

Definition 12. Bayesian minimum MAE estimator: $\hat{\theta} = \text{ median } [\theta|y]$ to minimize loss $L\left(\theta,\hat{\theta}\right) = \|\theta - \hat{\theta}\|_1$.

Definition 13. Gaussian penalty function: $\min_w \log p\left(y|w^T x\right) + \lambda\|w\|^2$.

Definition 14. Laplacian penalty function: $\min_w \log p\left(y|w^T x\right) + \lambda\|w\|_1$.

Definition 15. Sparsity penalty function: $\min_w \log p\left(y|w^T x\right) + \lambda\|w\|_0$.

Definition 16. Minimax Optimal Estimator: $\hat{\theta} = \arg\min_{\hat{\theta}} \sup_\theta R\left(\hat{\theta},\theta\right)$.

1.3 Distributions

Definition 17. Multivariate Normal Distribution: $p(x) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(\frac{-1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$; equivalently, $\log p(x) \propto \log|\Sigma| - \frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)$.

Definition 18. Binomial Distribution: $p(x) = \binom{n}{x} p^x (1-p)^{n-x}$.

Definition 19. Hypergeometric Distribution: $p(x) = \frac{\binom{b}{x}\binom{N-b}{n-x}}{\binom{N}{n}}$.

Definition 20. Multinomial Distribution: $p(x) = \binom{n}{x_1 x_2 \ldots x_k} \prod_{i=1}^{k} p_i^{x_i}$.

Definition 21. Exponential Distribution: $p(x) = \lambda e^{-\lambda x}$.

Definition 22. Gamma Distribution: $p(x) = \dfrac{x^{a-1} e^{-\frac{x}{b}}}{\Gamma(a) b^a}$.

Definition 23. Beta Distribution: $p(x) = \dfrac{x^{a-1}(1-x)^{b-1}}{Beta(a, b)}$, where Beta(a, b) $= \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

Definition 24. Exponential Family: $p(y|\theta) = b(y) \exp\left(\theta^T T(y) - \alpha(\theta)\right), \theta$ is the natural parameter and $T(y)$ is the sufficient statistic. Canonical form is when $T(y) = y$, and $\log p(y|\theta) = \sum_{i=1}^{n} \left(w^T x_i y_i - \alpha\left(w^T x_i\right)\right) + \log b(y_i)$.

1.4 Other Statistics, Algebra etc.

Definition 25. Kullback-Leibler Divergence: $D(p_1 \| p_0) = \mathbb{E}_{1[\Lambda(X)]} = \int p_1(x) \log \dfrac{p_1(x)}{p_0(x)} dx$.

Definition 26. Mahanalobis Distance: (x - $\mu)^T \Sigma^{-1} (x - \mu)$.

Definition 27. Sufficiency: $t(X)$ is sufficient if $p(x|t, \theta) = p(x|t)$.

Definition 28. Rao-Blackwellization: If f is an estimator and t is a sufficient statistic, then $\mathbb{E}[f(X) | t(X)]$ is the improved Rao-Blackwell estimator (in terms of MSE).

Definition 29. Characteristic Equation of X is $\det(\lambda I - X) = 0$, where λ are the eigenvalues.

Definition 30. Binomial Conjugate prior: Binomial(n, p) + Beta(α, β) = Beta$\left(\alpha + \sum_{i=1}^{n} x_i, \beta + n - \sum_{i=1}^{n} x_i\right)$, and Neg Binomal$(r, p)$ + Beta(α, β) = Beta$\left(\alpha + \sum_{i=1}^{n} x_i, \beta + rn\right)$.

Definition 31. Possion Conjugate prior: Poisson$(\lambda) + \Gamma(k, \beta)$ = Neg Binomial$\left(k + \sum_{i=1}^{n} x_i, \beta + n\right)$.

Definition 32. Normal Conjugate prior: Normal(μ, σ) + Normal(μ_0, σ_0) = Normal$\left(\dfrac{1}{\dfrac{n}{\sigma^2} + \dfrac{1}{\sigma_0^2}} \left(\dfrac{n}{\sigma^2}\bar{x} + \dfrac{\mu}{\sigma_0^2}\right), \dfrac{1}{\dfrac{n}{\sigma^2} + \dfrac{1}{\sigma_0^2}}\right)$ and Normal(μ, Σ) + Normal(μ_0, Σ_0) = Normal$\left(\left(\Sigma_{0^{-1}} + n\Sigma^{-1}\right)^{-1} \left(\Sigma_{0^{-1}}\mu_0 + n\Sigma^{-1}\bar{x}\right), \left(\Sigma_{0^{-1}} + n\Sigma^{-1}\right)^{-1}\right)$.

Definition 33. Uniform Conjugate prior: Uniform$(0, \theta)$ + Pareto(x_m, k) = Pareto$(\max\{x_1, ..., x_n, x_m\}, k + n)$.

Definition 34. Gamma Conjugate prior: Gamma (α, β) + Gamma (α_0, β_0) = Gamma $\left(\alpha_0 + n\alpha, \beta_0 + \sum_{i=1}^{n} x_i\right)$.

2 Theorems

2.1 Inequalities, Bounds etc.

Theorem 1. *Cauchy-Schwarz Inequality:* $|\mathbb{E}[XY]| \leqslant \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$.

Theorem 2. *Holder's Inequality: For* $\dfrac{1}{p} + \dfrac{1}{q} = 1, \mathbb{E}[|XY|] \leqslant (E[|X^p|])^{\frac{1}{p}} (E[|X^q|])^{\frac{1}{q}}$.

Theorem 3. *Markov's Inequality: For* $X \geqslant 0$ *and* $a > 0, \mathbb{P}\{X > a\} \leqslant \dfrac{\mathbb{E}[X]}{a}$.

Theorem 4. *Chebyshev's Inequality: For* $t > 0, \mathbb{P}\{|X - \mathbb{E}[X]| \geqslant t\} \leqslant \dfrac{Var[X]}{t^2}$.

Theorem 5. *Chebyshev-Cantelli Inequality: For* $t \geqslant 0, \mathbb{P}\{X - \mathbb{E}[X] > t\} \leqslant \dfrac{Var[X]}{Var[X] + t^2}$.

Theorem 6. *Jensen's Inequality: if f is convex, $\lambda f(x) + (1-\lambda) f(y) \geqslant f(\lambda x + (1-\lambda) y)$, then $\mathbb{E}[f(X)] \geqslant f(\mathbb{E}[X])$.*

Theorem 7. *Association Inequality: if f and g are increasing, $\mathbb{E}[f(X) g(X)] \geqslant \mathbb{E}[f(X)] \mathbb{E}[g(X)]$, and if f is increasing, g is decreasing, $\mathbb{E}[f(X) g(X)] \leqslant \mathbb{E}[f(X)] \mathbb{E}[g(X)]$.*

Lemma 1. *Fourth Moment: $\mathbb{E}[|X|] \leqslant \left(\mathbb{E}\left[X^2\right]\right)^{1.5} \left(\mathbb{E}\left[X^4\right]\right)^{-0.5}$.*

Lemma 2. *Chernoff bound $(1-\delta)$ confidence intervals for mean of $x_i \in [0,1]$ in k dimensions: $\pm\sqrt{\frac{\log\left(2k\delta^{-1}\right)}{2n}}$, and for standard deviation: $\sigma = \frac{1}{\sqrt{n}}$. The minimum number of data to ensure ε error with δ probability is $n \geqslant \frac{1}{2\varepsilon^2} \log\left(\frac{2k}{\delta}\right)$.*

Lemma 3. *Popoviciu's Inequality: If $\mathbb{P}\{m \leqslant z \leqslant M\} = 1$, then Var[Z] $\leqslant \frac{1}{2}(M-m)^2$.*

Theorem 8. *Hoeffding's Inequality: $X_i \in [a_i, b_i]$, $S_n = \sum_{i=1}^{n} X_i$, for each $t > 0$, $\mathbb{P}\{|S_n - E[S_n]| \geqslant t\} \leqslant 2\exp\left(-2t^2 \left(\sum_{i=1}^{n} (b_i - a_i)^2\right)^{-1}\right)$.*

Corollary 1. *Corollary to Hoeffding's Inequality: $X_i \in [a,b]$, $c = (b-a)^2$, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$, then $\mathbb{P}\{|\hat{\mu} - \mu| \geqslant t\} \leqslant 2\exp\left(-\frac{2nt^2}{c}\right)$.*

Lemma 4. *Lemma to proof Hoeffding's Inequality: $\mathbb{E}[Z] = 0, Z \in [a,b]$, then $\mathbb{E}\left[e^{sZ}\right] \leqslant \exp\left(\frac{1}{8}\left(s^2 (b-1)^2\right)\right)$.*

Theorem 9. *Sub Gaussian Tail Bound: if $\{Z_i\}_{i=1}^{n}$ are independent and $\mathbb{P}\{|Z_i - \mathbb{E}[Z_i]| \geqslant t\} \leqslant ae^{-b\frac{t^2}{2}}$, then $\mathbb{P}\left\{\frac{1}{n}\sum_i Z_i - \mathbb{E}[Z] > \varepsilon\right\} \leqslant e^{-cn\varepsilon^2}$ and $\mathbb{P}\left\{\mathbb{E}[Z] - \frac{1}{n}\sum_i Z_i > \varepsilon\right\} \leqslant e^{-cn\varepsilon^2}$ with $c = \frac{b}{16a}$.*

2.2 Linear Algebra

Theorem 10. *Singular Value Decomposition: $A = U\Sigma V^T$ satisfy $Av_i = \sigma_i u_i, A^T u_i = \sigma_i v_i$.*

Theorem 11. *Schur Complements:* $\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} I & 0 \\ CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{bmatrix} \begin{bmatrix} I & A^{-1}B \\ 0 & I \end{bmatrix} = \begin{bmatrix} I & BD^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{bmatrix} \begin{bmatrix} I & 0 \\ D^{-1}C & I \end{bmatrix}$.

Theorem 12. *Matrix Inversion Lemma: $(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B\left(D - CA^{-1}B\right)^{-1} CA^{-1}$.*

Theorem 13. *Sherman-Morrison Formula: $\left(A^{-1}uv^T\right)^{-1} = A^{-1} - \frac{A^{-1}uv^T A^{-1}}{1 + v^T A^{-1} u}$.*

Lemma 5. *Vector derivatives: $\frac{dc^T x}{dx} = c, \frac{dx^T x}{dx} = 2x, \frac{dx^T Ax}{dx} = \left(A + A^T\right) x$.*

2.3 Statistics

Theorem 14. *weak Law of Large Numbers: $\mathbb{E}[|X_i|] < \infty \Rightarrow \frac{1}{n}\sum_{i=1}^{n} X_i \rightarrow^p \mathbb{E}[X_i]$.*

Theorem 15. *strong Law of Large Numbers: $\mathbb{E}[|X_i|] < \infty \Rightarrow \frac{1}{n}\sum_{i=1}^{n} X_i \rightarrow \mathbb{E}[X_i]$ as .*

Theorem 16. *Central Limit Theorem: $\mathbb{E}[Z_i] = 0$, Var[Z_i] $= \sigma^2, \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) \rightarrow^d N\left(0, \sigma^2\right)$.*

Theorem 17. *Fisher-Neyman Factorization: $t(X)$ is sufficient iff $p(x|\theta) = a(x)\, b(t, \theta)$.*

Theorem 18. *Rao-Blackwell Theorem: let $t(X)$ be a sufficient statistic and define $g(t(X)) = \mathbb{E}[f(X)|t(X)]$, then $\mathbb{E}\left[(g(t(X)) - \theta)^2\right] \leqslant \mathbb{E}\left[(f(X) - \theta)^2\right]$, equal iff $f(X) = g(t(X))$.*

Lemma 6. *Convergence of Log-Likelihood to KL: $\hat{\theta}_n = \arg\max_\theta p(x|\theta) = \arg\min_\theta \sum_{i=1}^{n} \log\frac{q(x_i)}{p(x_i|\theta)} \rightarrow \arg\min_\theta D(q\|p_\theta)$.*

Theorem 19. *Asyptotic Distribution of MLE: Let $\hat{\theta}_n = \arg\max_\theta p(x|\theta)$, and $\mathbb{E}\left[\frac{\partial \log p(x|\theta)}{\partial\theta}\right] = 0$, then $\hat{\theta}_n \overset{asymp}{\sim} N\left(\theta, n^{-1} I^{-1}(\theta^\star)\right)$ where $[I(\theta^\star)]_{j,k} = -\mathbb{E}\left[\left.\frac{\partial \log p(x|\theta)}{\partial\theta_j \partial\theta_k}\right|_{\theta=\theta^\star}\right]$.*

Lemma 7. *KL-Divergence Information Matrix identity: if* $x|\theta \sim N(\theta, \sigma)$*, then* $\left.\frac{\partial^2 D\left(p\left(x|\theta\right) \| p\left(x|\theta^\star\right)\right)}{\partial \theta^2}\right|_{\theta=\theta^\star} = I\left(\theta^\star\right)$.

Theorem 20. *Gauss-Markov Theorem:* $\begin{bmatrix} x \\ y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}\right)$*, then* $y|x \sim N\left(\mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}\left(x - \mu_x\right), \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\right)$.

Theorem 21. *Direct Observation Model:* $y = W + \varepsilon, \varepsilon \sim N\left(0, \sigma^2 I\right)$*, and the soft-thresholding estimator* $\hat{w}_i = sign\left(y_i\right) \max\left\{|y_i| - \lambda, 0\right\}$*, oracle estimator* $\hat{w}_i = y_i \mathbb{1}_{\{|w_i|^2 \geqslant \sigma^2\}}$*, with* $\lambda = \sqrt{2\sigma^2 \log n}$*, then* $\mathbb{E}\left[\|\hat{w} - w\|^2\right] \leqslant (2\log n + 1)\left(\sigma^2 + \sum_{i=1}^{n} \min\left\{|w_i|^2, \sigma^2\right\}\right)$.

2.4 Optimization

Lemma 8. *Stepsize choice: if* $v_t = w_t - w^\star = \left(I - \gamma X^T X\right) v_1$*, then* $v_t \to 0$ *if the eigenvalues of (I -* $\gamma X^T X) < 1 \Rightarrow \gamma < \frac{2}{\lambda \max\left(X^T X\right)}$

Theorem 22. *Constant stepsize: If* $\|\nabla f_t(w)\| \leqslant G$ *and* $w^\star = \arg\min_w \sum_{t=1}^{T} f_t(w)$*, then gradient descent with* $\gamma_t = \gamma$ *starting at* w_1 *satisfies:* $\frac{1}{T}\sum_{t=1}^{T}\left(f_t\left(w_t\right) - f_t\left(w^\star\right)\right) \leqslant \frac{\|w_1 - w^\star\|^2}{2\gamma T} + \frac{\gamma}{2}G^2$.

Theorem 23. *Diminishing stepsize: If* $\|\nabla f_t(w)\| \leqslant G, \|w^\star\| \leqslant B$ *and* $w^\star = \arg\min_w \sum_{t=1}^{T} f_t(w)$*, then gradient descent with* $\gamma_t = \frac{1}{\sqrt{t}}$ *starting at* w_1 *satisfies:* $\frac{1}{T}\sum_{t=1}^{T}\left(f_t\left(w_t\right) - f_t\left(w^\star\right)\right) \leqslant \frac{2B^2 + G^2}{\sqrt{T}}$.

Lemma 9. *The inequality holds:* $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leqslant 2\sqrt{T}$.

2.5 Other results, formulas etc.

Lemma 10. *Empirical Classifier Error:* $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\hat{y}_i \neq y_i\}}$ *with mean* $\mathbb{E}\left[\hat{p}\right] = p$ *and Var*$[\hat{p}] = \frac{p(1-p)}{n}$ *, where* $p = \mathbb{P}\left\{\hat{y} \neq y\right\} = \mathbb{E}\left[\mathbb{1}_{\{\hat{y} \neq y\}}\right]$ *is the actual classifier error.*

Lemma 11. *KL-Divergence of Normal Distribution: With same varaince:* $X|Y \sim N\left(\mu_j, \Sigma\right)$ *with common covariance is* $D\left(p_0 \| p_1\right) = D\left(p_1 \| p_0\right) = \frac{1}{2}\left(\mu_1 - \mu_0\right)^T \Sigma^{-1}\left(\mu_1 - \mu_0\right)$*. With different variances:* $D\left(N\left(\mu_0, \Sigma_0\right) \| N\left(\mu_1, \Sigma_1\right)\right)$ *is* $\frac{1}{2} tr\left(\Sigma_1^{-1}\Sigma_0\right) + \frac{1}{2}\left(\mu_1 - \mu_0\right)^T \Sigma_1^{-1}\left(\mu_1 - \mu_0\right) - d + \log\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right)$.

Lemma 12. *Optimal Bayes binary Classifier with common covariances and equal prior is* $\hat{y}(x) = 1$ *if* $2\left(\mu_1 - \mu_0\right)^T \Sigma^{-1} x \geqslant \mu_0^T \Sigma \mu_0 - \mu_1^T \Sigma \mu_1$ *is linear in* x.

Lemma 13. *Non-negative Expected Value: If* $Y \geqslant 0$*, then* $\mathbb{E}[Y] = \sum_{i=1}^{\infty} \mathbb{P}\{Y \geqslant i\}$.

Lemma 14. *Sum formulas:* $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}; \sum_{i=1}^{n} i^2 = \frac{n(n+1)(n+2)}{6}$.

Lemma 15. *Bayesian Linear Regression with prior* $w \sim N\left(0, \sigma_w^2 I\right)$ *has* $\hat{w} = \left(X^T X + \lambda I\right)^{-1} X^T y$*, where* $\lambda = \frac{\sigma^2}{\sigma_w^2}$.

Lemma 16. *Minimax Optimal Estimator: if* $\hat{\theta}_p = \arg\min_{\hat{\theta}} \int R\left(\hat{\theta}, \theta\right) p(\theta)\, d\theta$*, and* $\int R\left(\hat{\theta}_p, \theta\right) p(\theta)\, d\theta = \sup_\theta R\left(\hat{\theta}_p, \theta\right)$*, then* $\hat{\theta}_p$ *is minimax optimal. In particular, if* $R\left(\hat{\theta}_p, \theta\right)$ *is constant, then it is minimax.*

Lemma 17. *Subgradients: for* $\|w\|_1$ *is sign*(w)*; for* $\max\left\{0, x^T w\right\}$ *is* $x \mathbb{1}_{x^T w > 0}$.