Midterm 2 note sheet

Table 1: Performance metrics.

Confusion matrix rows: the true class	Y
Confusion matrix columns: the predicted class	$\hat{Y}$
True positives	T_p
False positives	F_p
False negatives	F_n
True negatives	T_n
Total	$N = T_p + F_p + F_n + T_n$
Accuracy	$\text{Acc} = (T_p + T_n)/N$
Precision	$P = \frac{T_p}{T_p + F_p}$
Recall	$R = \frac{T_p}{T_p + F_n}$
Specificity	$\text{TNR} = \frac{T_n}{F_p + T_n}$
False positive rate	$\text{FPR} = \frac{F_p}{F_p + T_n}$
F_1 score	$F_1 = \frac{2PR}{P+R}$

Table 2: Logistic regression.

Logistic function	$\text{expit}(z) = \frac{1}{1+\mathrm{e}^{-z}}$
Value at zero	$\text{expit}(0) = \frac{1}{2}$
Symmetry	$\text{expit}(-z) = 1 - \text{expit}(z)$
Derivative	$\text{expit}'(z) = \text{expit}(z)(1 - \text{expit}(z))$
Logistic regression model	$\Pr(Y = 1 \mid \underset{\sim}{x}) = \text{expit}\left(\vec{\beta} \cdot \underset{\sim}{x}\right)$
Odds	$\omega = \pi/(1-\pi)$
Log odds (logit)	$\eta = \ln \frac{\pi}{1-\pi}$
Probability to log odds	$\pi \underset{\omega/(1+\omega)}{\overset{\pi/(1-\pi)}{\rightleftarrows}} \omega \underset{\mathrm{e}^{\eta}}{\overset{\ln \omega}{\rightleftarrows}} \eta$
Logistic regression model, log-odds form	$\eta = \vec{\beta} \cdot \underset{\sim}{x}$
Odds multiplier of β_j	e^{β_j}
Threshold rule	$h(\underset{\sim}{x}) = \mathbb{1}(\Pr(Y = 1 \mid \underset{\sim}{x}) > \theta)$
Decision boundary at a threshold	$\vec{\beta} \cdot \underset{\sim}{x} = \ln \frac{\theta}{1-\theta}$
Probability of one Bernoulli outcome	$\Pr(Y = y \mid \underset{\sim}{x}) = \pi^y (1-\pi)^{1-y}$
Log loss	$L(y, \pi) = -y \ln \pi - (1-y) \ln(1-\pi)$
Total log loss	$C(\vec{\beta}) = \sum_{i=1}^{n} L(y_i, \pi_i)$
Gradient of the total log loss	$\nabla C(\vec{\beta}) = \sum_{i=1}^{n} (\pi_i - y_i)\, \underset{\sim}{x}_i$
Hessian of the total log loss	$\nabla^2 C(\vec{\beta}) = \sum_{i=1}^{n} \pi_i (1-\pi_i)\, \underset{\sim}{x}_i \underset{\sim}{x}_i^{\top}$
Newton update	$\vec{\beta}_t = \vec{\beta}_{t-1} - \left(\nabla^2 C(\vec{\beta}_{t-1})\right)^{-1} \nabla C(\vec{\beta}_{t-1})$

Table 3: Gradient descent.

Power rule	$\frac{dw^k}{dw} = kw^{k-1}$
Gradient	$\nabla C(\vec{\beta}) = \left[\frac{\partial C}{\partial \beta_1}, \ldots, \frac{\partial C}{\partial \beta_p}\right]^{\top}$
Step size	$h > 0$
Gradient descent update	$\vec{\beta}_t = \vec{\beta}_{t-1} - h \nabla C(\vec{\beta}_{t-1})$

Table 4: Logarithms and exponentials.

Arguments of the logarithm	$a, b > 0$
Log of one	$\ln 1 = 0$
Exponential of zero	$\mathrm{e}^0 = 1$
Log of an exponential	$\ln(\mathrm{e}^s) = s$
Exponential of a log	$\mathrm{e}^{\ln a} = a$
Log of a reciprocal	$\ln(1/a) = -\ln a$
Exponential of a negative	$\mathrm{e}^{-s} = 1/\mathrm{e}^s$
Log of a product	$\ln(ab) = \ln a + \ln b$
Log of a quotient	$\ln(a/b) = \ln a - \ln b$
Log of a power	$\ln(a^k) = k \ln a$
Exponential of a sum	$\mathrm{e}^{s+t} = \mathrm{e}^s \mathrm{e}^t$
Exponential of a difference	$\mathrm{e}^{s-t} = \mathrm{e}^s / \mathrm{e}^t$
Power of an exponential	$(\mathrm{e}^s)^k = \mathrm{e}^{ks}$

Table 5: Decimal values of exponentials and logarithms.

	e	e^{-1}	$\ln 2$	$\ln 3$	$\ln 4$	$\ln 5$	$\ln 10$
Value	2.718	0.368	0.693	1.099	1.386	1.609	2.303

Table 6: Generalization.

Outcome	$Y = \mu(x) + \varepsilon$
Noise variance	$\mathrm{Var}(\varepsilon) = \sigma^2$
Prediction error	$\varepsilon_{\text{pred}} = (\hat{y} - Y)$
Bias	$\mathrm{Bias}(\hat{y}) = \mathbb{E}\left[\varepsilon_{\text{pred}}\right]$
Deviation of the prediction	$\mathbb{d}_{\hat{y}} = \hat{y} - \mathbb{E}_{\mathcal{D}_{\text{train}}}[\hat{y}]$
Variance	$\mathrm{Var}(\hat{y}) = \mathbb{E}\left[\left(\mathbb{d}_{\hat{y}}\right)^2\right]$
Expected squared error	$\mathrm{ESE}(x) = \mathbb{E}\left[\left(\varepsilon_{\text{pred}}\right)^2\right]$
Expected squared error decomposition	$\mathrm{ESE}(x) = \left(\mathrm{Bias}(\hat{y})\right)^2 + \mathrm{Var}(\hat{y}) + \sigma^2$
Empirical risk	$\hat{R}(h; \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, h(\underset{\sim}{x}_i))$
Training error	$\hat{R}(h; \mathcal{D}_{\text{train}})$
Test error	$\hat{R}(h; \mathcal{D}_{\text{test}})$
Cross-validation estimate, equal-size-fold case	$\hat{R}_{k\text{-fold}} = \frac{1}{k}\sum_{j=1}^{k} \mathrm{MSE}_j$

Table 7: Linear regression.

Risk	$R(\vec{\beta}) = \mathbb{E}\left[(\vec{\beta} \cdot \underset{\sim}{D}(\underset{\sim}{x}) - y)^2\right]$
Sample MSE	$\mathrm{sMSE}\left(\vec{\beta}\right) = \frac{1}{n}\sum_{i=1}^{n} (\vec{\beta} \cdot \underset{\sim}{D}(\underset{\sim}{x}_i) - y_i)^2$
Sample MSE, matrix form	$\mathrm{sMSE}\left(\vec{\beta}\right) = \frac{1}{n}\left\|\mathbf{D}\vec{\beta} - \underset{\sim}{y}\right\|_2^2$
Gradient of the sample MSE	$\nabla \mathrm{sMSE}\left(\vec{\beta}\right) = \frac{2}{n}\mathbf{D}^\top(\mathbf{D}\vec{\beta} - \underset{\sim}{y})$
Normal equations	$\mathbf{D}^\top\mathbf{D}\hat{\vec{\beta}} = \mathbf{D}^\top \underset{\sim}{y}$
Unique solution condition	$\det(\mathbf{D}^\top\mathbf{D}) \neq \mathbb{0}$
Solution	$\hat{\vec{\beta}} = (\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{D}^\top \underset{\sim}{y}$
Fitted values	$\underset{\sim}{\hat{y}} = \mathbf{D}\hat{\vec{\beta}}$
Residual vector	$\underset{\sim}{\hat{\varepsilon}} = \underset{\sim}{y} - \underset{\sim}{\hat{y}}$

Table 8: Design matrix.

Intercept trick	$\underset{\sim}{D}(\underset{\sim}{x}) = \left[1, x_1, \ldots, x_p\right]^\top$
Predictor matrix	$\mathbf{x} = \left[\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n\right]^\top \in \mathbb{R}^{n \times p}$
Design function	$\mathbf{D}(\mathbf{x}) = \left[\underset{\sim}{D}(\underset{\sim}{x}_1), \ldots, \underset{\sim}{D}(\underset{\sim}{x}_n)\right]^\top$
Design matrix	$\mathbf{D} = \mathbf{D}(\mathbf{x})$
Design matrix shape	$\mathbf{D} \in \mathbb{R}^{n \times p^*}$
Coefficient vector shape	$\vec{\beta} \in \mathbb{R}^{p^*}$
Predictions	$\underset{\sim}{\hat{y}} = \mathbf{D}\vec{\beta}$

Table 9: Feature expansion.

Feature map	$\underset{\sim}{D} : \mathbb{R}^p \to \mathbb{R}^{p^*}$
Linear model in the features	$\underset{\sim}{h}(\underset{\sim}{x}) = \vec{\beta} \cdot \underset{\sim}{D}(\underset{\sim}{x})$
Monomial	$\underset{\sim}{m}(\underset{\sim}{x}) = \prod_{j=1}^{p} \underset{\sim}{x}_j^{a_j}$
Monomial degree bound	$\sum_{j=1}^{p} a_j \leq d$

Table 10: Vector calculus.

Linear form	$\frac{\partial}{\partial \vec{\beta}} \underset{\sim}{c} \cdot \vec{\beta} = \underset{\sim}{c}$
Squared norm	$\frac{\partial}{\partial \vec{\beta}} \vec{\beta} \cdot \vec{\beta} = 2\vec{\beta}$
Matrix-vector product, constant $\mathbf{A}$	$\frac{\partial}{\partial \vec{\beta}}(\mathbf{A}\underset{\sim}{v}) = (\frac{\partial}{\partial \vec{\beta}}\underset{\sim}{v})\mathbf{A}^\top$
Special case $\underset{\sim}{v} = \vec{\beta}$	$\frac{\partial}{\partial \vec{\beta}}(\mathbf{A}\vec{\beta}) = \mathbf{A}^\top$
Next rule needs $\mathbf{S}$ symmetric	$\mathbf{S} = \mathbf{S}^\top$
Quadratic form	$\frac{\partial}{\partial \vec{\beta}} \vec{\beta}^\top \mathbf{S} \vec{\beta} = 2\mathbf{S}\vec{\beta}$

Table 11: Shapes and calculus.

Matrix product shape	$(a \times b)(b \times c) \to a \times c$
Transpose shape	$a \times b \to b \times a$
Dot product	$\underset{\sim}{x} \cdot \underset{\sim}{y} = \sum_{i=1}^{n} x_i y_i$
Sample mean	$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
Sum rule	$\frac{d}{dw}\sum_{i=1}^{n} u_i = \sum_{i=1}^{n} \frac{du_i}{dw}$
Chain rule	$\frac{dh(u)}{dw} = \frac{dh(u)}{du}\frac{du}{dw}$
Minimum: first derivative is zero	$\frac{\partial \hat{R}}{\partial w} = 0$
Minimum: second derivative is positive	$\frac{\partial^2 \hat{R}}{\partial w^2} > 0$