ST3248 Midterm Cheatsheet

by Devansh Shah (adapted from Lingjie's summary notes)

Definitions

Supervised	:= Predict an outcome variable Y based on one or more inputs.
Unsupervised	:= No supervising outputs, learns rs between variables.
Error	:= $Y - \hat{Y}$
Bias	:= $E(\hat{\theta} - \theta)$, inflexible model has higher bias
Variance	:= highly flexible method follow data closely
Bias-variance trade-off	:= U shaped test MSE curve, higher flexibility result in low bias but high var
Curse of dimensionality	:= poor model performance at higher dimension
Parametric model	:= model with pre-determined specification
Non-parametric model	:= model without any pre-determined specification
Over-fitting	:= model is tailored to training data and perform badly on test data
Under-fitting	:= model failed to capture the underlying relationship in data
Reducible error	:= error can be reduced with better method or more data
Irreducible error	:= random error ϵ unable to be predict from X

Parametric vs. Non-Parametric

- Parametric approach will work best when true f is similar to parametric form chosen
- However, if model specification is wrong, non-parametric outperform parametric in most cases
- Some methods (e.g. KNN) perform worse than parametric methods in high dimension (curse of dimensionality)

Regression problems

Key problem: estimate

$$E(Y|X = x) = E(f(x) + \varepsilon|X = x) = f(x)$$

Irreducible error, Bias Var trade-off

Mean Squared Error
= reducible + irreducible error
= $bias^2 + var$ + irreducible error

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

where $Y = f(x) + \varepsilon$
Assumptions of linear regression:

- $\varepsilon \perp X$ (constant variance aka homoscedasticity)
- $E[\varepsilon] = 0$

For simple linear regression,

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x}$$

Mean squared error

$$L(e) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

Standard Errors

Here, $\sigma^2 = Var(\varepsilon)$ and we assume that the errors ε_i are uncorrelated with common variance. In practice, we can estimate σ by the residual standard error (RSS).

$$SE(\hat{\beta}_0)^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$
$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

t-statistic for hypothesis testing: $\frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)}$). Under H_0, t has t-distribution with $n - 2$ degrees of freedom.

For a linear regression model with p predictors, the F-statistic:

$$F = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)} \sim \chi^2_{p,n-p-1}$$

Residual standard error, R^2, Adjusted R^2

$$RSE = \sqrt{\frac{1}{n - p - 1}RSS}$$
$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$
$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$
$$AdjR^2 = 1 - \frac{RSS/(n - p - 1)}{TSS/(n - 1)}$$

- R^2 can be interpreted as the "proportion of variance in y that can be explained by the model".
- For simple linear regression, $R^2 = r^2$ where r is the sample correlation, given by:

$$r = \hat{Cor}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Higher adjusted $R^2 \implies$ better model.

Model Selection

- Forward Selection - start with null model and greedily add the variable that results in lowest RSS.
- Backward Deletion - start with the full model and greedily remove the variable with the largest p-value.
- Mixed Selection - Interleave forward and backward selection with a fixed threshold p-value until all variables in the model have p-values smaller than threshold, and any variable added to the model will have p-value larger than the threshold.

Confidence and Prediction Intervals

Confidence Interval - Used when we're estimating how close $\hat{Y}$ will be to $f(X)$. Since we're estimating the average value of the response, we can ignore the random error ε since it's average is zero.

Prediction Interval - Used when we're making a prediction for an individual Y given X. This needs to account for the random error ε

So, prediction interval is always wider than a confidence interval (for the same degree of confidence).

Remarks on Regression

- We interpret β_j as the average effect on Y of one unit increase in X_j, *holding all other predictors fixed.*
- p-value for each individual predictor provides information about whether an individual predictor is related to the response, *after adjusting for the other predictors.*
- Even if one of the p-values is small, it does NOT mean that the overall F-statistic must have a small p-value.

Classification

Key problem: estimate

$$p_j(x) = P(Y = j|X = x_0)$$

Irreducible error

Bayes error rate

$$1 - E(\max_j P(Y = j|X))$$

lowest possible test error rate (irreducible error)

Bayes classifier

choose $\hat{y}_0 = j$ from data s.t.

$$\max_j P(Y = j|X = x_0)$$

Bayes decision boundary: $P(Y = 1|X) = P(Y = 2|X) = 0.5$
Bayes classifier achieves Bayes error rate. However, we need to estimate $P(Y = j|X = x_0)$ before using Bayes classifier (which is not possible in practice).

Error rate

$$L(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

Logistic Regression

The logistic function ensures that the output is in the range $[0, 1]$ for *all* values of X:

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

For the "simple" logistic regression (only 1 predictor), the decision boundary is linear. But we can add more predictors to increase the flexibility of the decision boundary.

Odds

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

log-odds/logit

$$log(\frac{p(X)}{1 - p(X)}) = \beta_0 + \beta_1 X$$

Likelihood function

$$L(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

Confusion matrix

True class (horizontal)

Predicted class		Positive	Negative
	Positive	TP	FP
	Negative	FN	TN

FP rate := % negative examples that are classified as positive
FN rate := % positive examples that are classified as negative
Sensitivity := True positive rate $(TP/(TP + FP))$
Specificity := True negative rate $(TN/(TN + FP))$

Linear Discriminant Analysis

- LDA is more stable than logistic regression (especially when the classes are well-separated)
- LDA is more popular when we have more than 2 response classes.

We define $\pi_k = P(Y = k)$ and $f_k(X) = Pr(X = x|Y = k)$, then

$$p_k(x) = P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^{K} \pi_i f_i(x)}$$

If we assume the conditional distributions $f_k(x)$ are normal with their own means μ_k but common variance σ^2, we end up with a linear boundary.
Then, the formula for LDA is equivalent to picking k with the largest value of:

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

Since $\delta_k(x$ is linear in x, the decision boundary is also linear.

In practice, we estimate the quantities using information from our sample:

$$\hat{\pi}_k = \frac{n_k}{n}$$

$$\hat{\mu}_k = \frac{1}{n_k}\sum_{i:y_i=k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n - K}\sum_{k=1}^{K}\sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

In case of multiple predictors, we assume that the $X = (X_1, X_2, \ldots, X_p)$ has a (conditional) multi-variate normal distribution with class-specific mean vector and common covariance matrix.

Quadratic Discriminant Analysis

QDA relaxes the common variance assumption: so the distribution of $X$for each class of Y is normally distributed and has its own mean *and covariance matrix*.
The discriminant function $\delta_k(x)$ has a term quadratic in x, hence the name. This results in a quadratic decision boundary.
QDA is more flexible than LDA. So, QDA has lower bias, higher variance (because more parameters to estimate).

AUC, ROC

ROC: receiver operating characteristic curve
AUC: area under the curve

- Compares FP rate (x-axis) and TP rate (y-axis)
- Higher AUC, larger area = better
- Useful to compare different probability threshold
- When threshold = 1, FP=0, TP=0 (lower left)
- When threshold = 0, FP=1, TP=1 (upper right)

Note: $\hat{P}(Y|X) \geq$ threshold will be classified as positive

Cross-validation (CV)

Key problem: estimate test MSE using given data set. Here, the "bias" and "variance" is in estimating the test MSE, *not the model parameters*.

Validation-set approach

Divide samples into:
1. Training set 2. Validation/hold-out set
Limitation

- High variance (since the model performance on the validation set depends heavily on the observations present from the training data set, which can change in each iteration)
- Overestimates test error as only subset of data is used in training (bias can possibly be reduced if we use complete data set)

K-fold Cross-validation

1. Randomly split n samples into K blocks, each block has $n_K = n/K$ obs
2. For model $j = 1, \cdots, M$:
   [2.1] For block $k = 1, \cdots, K$
   [2.1.1] Fit model j on all blocks except block k
   [2.1.2] Evaluate model j based on block k
   [2.2.2] Calculate $CV_{(K)}^{(j)} = \sum_{k=1}^{K}(n_k/n)MSE_k$
   Note $(n_k/n) \approx 1/K$

Estimated test error

$$CV_{(K)} = \frac{1}{K}\sum_{i=1}^{K} MSE_i$$

Variance

$$\hat{Var}(CV_{(K)}) = \sum_{k=1}^{K}\frac{(MSE_k - \bar{MSE}_k)^2}{K - 1}$$

$$= \frac{1}{K}\hat{Var}(CV_k) + 2\sum_{i<j}^{K} Cov(CV_i, CV_j)$$

Limitation

- Bias still exist (min bias when $K = n$)
- High variance due to overlapping blocks used for training model in each iteration
- CV should be performed before feature selection step

Leave-one out CV (LOOCV)

Using Cross-validation with $K = n$
Special result for least-squares linear, polynomial regression

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$

$h_i :=$ leverage, $n :=$ number of samples (i.e. LOOCV)
Limitation

- High variance since estimates from each fold are highly correlated (nearly identical training data)
- Computational intensive (in general case)

Bootstrap

Key idea: resampling with replacement
Suppose we're estimating the standard error and bias of the statistic $\hat{\alpha}$ (obtained using original data),

1. For bootstrap sample $r = 1, \cdots, B$:
   Randomly select n observations with replacement from the training set.
   Compute $\hat{\alpha}^{*r}$ for this sample.

Then, define $\hat{\alpha}' = \frac{1}{B}\sum_{r=1}^{B}\hat{\alpha}^{*r}$ (mean of all statistics across all bootstrap samples)

$$\hat{SE}(\alpha) = \sqrt{\frac{1}{B - 1}\sum_{r=1}^{B}(\hat{\alpha}^{*r} - \hat{\alpha}')^2}$$

$$\hat{Bias}(\hat{\alpha}) = \frac{1}{B}\sum_{r=1}^{B}\hat{\alpha}^{*r} - \hat{\alpha}$$