1 Linear Regression

- We assume the model as: $y = XB + \epsilon$ (matrix form) - We see this holds for a single case, as the ith observation of $XB + \epsilon = x_1\beta_1 + x_2\beta_2 + \ldots + \epsilon_i$ - Assume ϵ distributed $MVN(0, \sigma^2 I)$ - Errors are centered around 0, or the expected value of the errors is 0 - They all have the same variance (identially distributed) - There is no covariance between the errors (independently distributed)

- Linear: We model the response as linear combination of the betas. Does *not* mean we can't have curves. We can have predictors like x^2, they don't need to be linear - Finds the line / hyperplane of best fit, where best is defined as minimizing the sum of squared residuals (SSR), defined as: $\sum(y_i - y_p red)^2$ - Squared so as to treat negative mistakes equally as positive mistakes - Can also minimize the average quantity (divide by n). These are equivalent - positive scalars do not affect argmin problems

- The standard solution is given by $\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y$ - Has the name Ordinary Least Squares (OLS) - X is the **design matrix**- rows denote observations, columns denote features - Obtained via: 1. Rewriting SSR in matrix form: $(y - X\beta)^T(y - X\beta)$ 2. Taking partial wrt β (matrix differentiations work similarly to scalars - we can work out the individual betas and see we get equivalent results) 3. Set equal to 0 -¿ solve for beta - Unbiased estimators - High bias - Not possible in high-dimensional setting - need to introduce regularization

- Model comparison: - R^2: Percentage of variance explained by model. Monotonically increases with more predictors (cannot use to compare models with different number of predictors) - Adjusted R^2: Like R^2, but "penalizes" on adding more predictors. Can compare models with different number of predictors - AIC / BIC

2 Logistic Regression

Designed for binary response ($y \in \{0, 1\}$)

Instantiation of generalized linear model:

1. Linear predictor (as before): XB (no epsilon)
2. Link function: Relates the conditional expectation ($\mu = E[Y|X]$) to linear predictor via g(mu) = XB. - For logistic regression, the link function is the logit (log odds)
3. Family of probability distribution in exponential family: Bernouili

No closed form solution

No residuals. Why? - Bernouili distribution does not have constant variance: parameters closer to 0.5 have larger variance

3 Lasso Regression

3.1 Why?

Classic linear regression does not work in high-dimensional setting (X^TX is singular, and therefore its inverse is not defined). Lasso is

Hope: Better performance on unseen data by utilizing the bias-variance tradeoff. We introduce more *bias* to the model (doesn't fit training data as well), in hopes that the subsequent decrease in variance will be worth it

- Depending on magnitude of penalization, will perform feature selection - How does it feature select? - Geometrically, the L1 penalty can be thought of as a diamond (2d), tetrahedron (3d), etc. shape around the origin. With the penalty term, we're saying the beta solutions need to live in this shape. So we radiate out in equal SSR contours from the OLS estimate until we hit this shape, and we call that our solution. As the dimensions go up, it's a virtual certainty that we will hit some sharp "corner" of this shape. This is because the curse of dimensionality pushes more and more volume into the corners in high dimensions, very spindly. And by definition, the corners of this shape are axes, and axes are defined such that some beta values are 0. 2 extremes

- $\lambda = 0$: No penalty on beta coefficients. Recover OLS estimates
- $\lambda \leftarrow \infty$: Extreme penalty on beta coefficients. Fits an intercept only hyperplane for the model

Will not penalize the intercept

Adds L1 penaliation term to beta coefficients

Can be framed either as minimizing a loss function, or a constrained optimization problem (equivalent specifications)

4 Ridge Regression

4.1 Why?

- See Lasso. Introduce bias for reduced variance (hopefully) L2 penalty on non-intercept beta coefficients

Will *not* perform feature selection (although beta values can get arbitrarily small)

Extremes as above: Can either recover OLS estimates with no penalty, or intercept only model with extreme penalty

5 Decision Tree

- Split based on entropy or gini impurity

6 Random Forest

2 key ideas

- Bagging ("forest"): Reduce the high variance of a single decision tree by taking many bootstrapped samples and averaging bootstrapped samples obtained by taking samples from the original dataset *with replacement* until we have one of the same size
- Random feature subsets ("random"): If we have one really important feature, then all bagged trees will use it at the top level -¿ trees will be correlated. Averaging correlated quantities does not reduce variance as much - Decorrelate trees by only allowing the trees to choose from a random subset of variables at each split

Commonly consider splits of $\sqrt{p}$ or $log_2(p)$

7 Naive Bayes

8 Support Vector Machine

9 XG Boost