• Because KNN is completely non parametric, we can expect this approach to dominate LDA and logistic regression when the boundary is highly non-linear, provided n is very large and p is small, but KNN doesn't tell us which predictors are important.

Generalized linear model: Poisson regression

when Y is neither qualitative nor quantitative → Poisson regression

Poisson distribution $Pr(Y=k) = \frac{e^{-\lambda}\lambda^k}{k!}$ for $k = 0, 1, 2 \ldots$ used to model counts.

$\lambda = E(Y) = Var(Y)$

we want the mean to vary as function of the covariates $\lambda = E(Y) = \lambda(X_1, \ldots X_p)$
$= e^{\beta_0 + \beta_1 X_1 + \ldots \beta_p X_p}$

to estimate $\beta_0, \ldots \beta_p$ we use the maximum likelihood $l(\beta_0, \ldots \beta_p) = \prod_{i=1}^{n} \frac{e^{-\lambda(x_i)}\lambda(x_i)^{y_i}}{y_i!}$

where $\lambda(x_i) = e^{\beta_0 + \beta_1 x_{i1} + \beta_p x_{ip}}$

Interpretation: an increase in X_j by one unit is associated with a change in $E(Y) = \lambda$ by a factor of $\exp(\beta_j)$

Generalized models in generality

Linear, logistic and poisson regression share common characteristic

→ Use $X_1, \ldots X_p$ to predict Y

→ conditionnal on $X_1, \ldots X_p$, Y belong to a family
- linear: gaussian
- logistic: Bernoulli
- Poisson: Poisson

for linear: $E(Y|X_1, \ldots X_p) = \beta_0 + \beta_1 X_1 + \ldots \beta_p X_p$

$\mu \Rightarrow$ logistic: " $= Pr(Y=1|X_1, \ldots X_p) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_p \ldots}}{1 + e^{\beta_0 + \beta_1 X_1 \ldots}}$

Poisson " $= \lambda(X_1, \ldots X_p) = e^{\beta_0 + \beta_1 X_1 + \ldots \beta_p X_p}$

can be expressed using link function n, which applies a transformation to $E(Y|X_1, \ldots X_p)$ so the transformed mean is linear

$n(\underbrace{E(Y|X_1, \ldots X_p)}_{\mu}) = \beta_0 + \beta_1 X_1 + \ldots \beta_p X_p$

- linear: $n(\mu) = \mu$
- logistic: $n(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$
- Poisson: $n(\mu) = \log(\mu)$

The gaussian, bernoulli, poison distribution are all members of a wider class of distribution, known as the exponential family (also include exponential, gamma, negative binomial). In general, we can perform a regression by modeling the response Y as coming from a particular member and the transforming the mean so that the transformed mean is linear function of the predictors.

Resampling methods *

- cross validation can be used to estimate the test error associated with a given statistical learning method in order to evaluate its performance, or to select the appropriate level of flexibility (model selection)
(model assessment)

Traing validation -> evaluate highly variable

- Leave-one-out cross validation (LOOCV)

$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$ -> less bias, no randomness in the training validation set split than CV

- k-fold: split in k

k=3 good one are k=5 or 10 -> good for low variance (compared to LOOCV)

- The Bootstrap is a widely applicable method that can be used to quantify the uncertainty associated with a given estimator or statistical learning method. Rather than repeatedly obtaining independent data set from the population, we instead obtain distinct data set by repeatedly sampling observation (with replacement) from the original dataset.