CSCI-GA.2565
Machine Learning
Homework 1

Michael Lukiman
mLL469

Problem 1 - Probability

(a) Given: • sample $\vec{x} = (x_1, \ldots, x_n)$
• Bernoulli distribution
$\longrightarrow$ Find: maximum likelihood estimator of p

Let's say probability of $(y=1|\vec{x}) \equiv \phi$

$\therefore \quad = \phi^k (1-\phi)^{(1-k)}$ where $k \in \{0, 1\}$ (belongs to; range)

$\therefore$ likelihood $L(\phi) = \prod_{i=1}^{n} \left[\phi^{x_i} (1-\phi)^{(1-x_i)} \right]$

$\therefore$ log likelihood $\ell(\phi) = \log \phi \sum_{i=1}^{n} x_i + \log(1-\phi) \sum_{i=1}^{n} (1-x_i)$

$\rightarrow \dfrac{\partial \ell(\phi)}{\partial \phi} = \Sigma x_i / \phi \; - \; \Sigma(1-x_i) / 1-\phi \overset{\text{set}}{=}$ zero

$\xrightarrow[\text{by } p(1-p)]{\text{multiply}} (\Sigma x_i) - \hat{\phi}(\Sigma x_i) = \hat{\phi} \Sigma (1-x_i) \longrightarrow \hat{\phi} n = \Sigma x_i$

$\therefore \hat{\phi} = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}$

since $\dfrac{\partial^2 \ell(p)}{\partial p^2}$ is negative, we confirm this is a maximization

(b) Now suppose the distribution is uniform

Find: MLE:

$$P(y_i|\vec{x};a,b) = \begin{cases} 0 & \text{for } x<a \\ 1/(b-a) & \text{for } a \le x \le b \\ 0 & \text{for } x > b \end{cases}$$

$\therefore$ likelihood $L(a,b|\vec{x}) = \prod_{i=1}^{n} \frac{1}{b-a}$

$\therefore \ell(\cdot) = \sum_{i=1}^{n} \log\left(\frac{1}{b-a}\right) = \sum_{i=1}^{n} \left(\log(1) - \log(b-a)\right)$

$\xrightarrow[a]{\text{partial to}} \frac{\partial \ell(\ldots)}{\partial a} = \sum_{i=1}^{n} \left(0 - \frac{-1}{b-a}\right) \therefore \hat{a} = \min(\vec{x})$

$\xrightarrow[b]{\text{partial to}} \frac{\partial \ell(\ldots)}{\partial b} = \sum_{i=1}^{n} \left(0 - \frac{1}{b-a}\right) \therefore \hat{b} = \max(\vec{x})$

$\because$ we want both to approach zero

(Both 2nd derivatives are negative) ✓

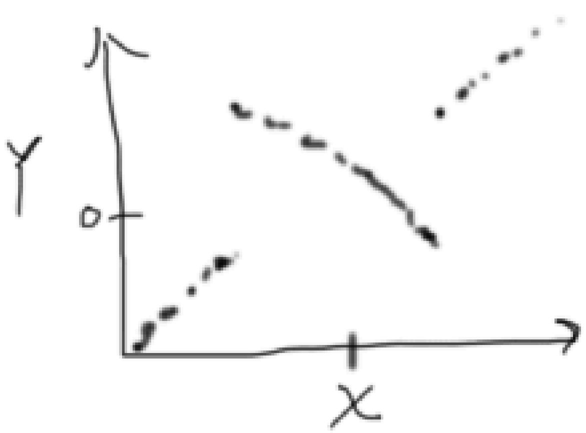
(c) Given: • X, Y random
• both marginally Gaussian
• correlation = 0

Find: Are they independent?

given $\not\Rightarrow$ independence

Counter example: X maps to $f(z) = z$

Y maps to $g(z) = \begin{cases} f(z) & \text{for } |z| \ge c \\ -f(z) & \text{for } |z| < c \end{cases}$



Y depends on X, yet the raw correlation equals zero by the average, by some c (not all).

X has a gaussian mean and variance, as does Y, by its CDF.

This all considered, given $\not\Rightarrow \neg$ independent.

$\therefore$ need more info.

Problem 2 - Poisson GLM's

Given:
- sample pairs $\vec{xy} = \{(x_1,y_1),\dots(x_n,y_n)\}$
- $x_i \in \mathbb{R}^p$ (p dimensions)
- $y_i \in \mathbb{N}\{0,1,2,3\dots\}$ for $i=1,\dots,n$ (countable)
- linear coefficients are $\vec{\theta}$, also then in $\mathbb{R}'$

(a) Find: log likelihood ℓ aka LL

For GLM form λ in Poisson as $e^{\vec{\theta}^T\vec{x}}$ (via η in exponential family form)

$$L(\vec{\theta}|\vec{xy}) = \prod_{i=1}^{n}\left[e^{y_i\vec{\theta}^T x_i}\, e^{-e^{\vec{\theta}^T x_i}} / y_i!\right] \quad \because p(y_i|x_i) = \frac{\lambda^{y_i}}{y_i}e^{-\lambda}$$

$$\xrightarrow{\text{take log}} LL = \sum_{i=1}^{n}\log(L(\vec{\theta}|\vec{xy})) = \sum\left(\log(e^{y_i\vec{\theta}^T x_i}\, e^{-e^{\theta^T x_i}}) - \log(y!)\right) = \boxed{\sum_{i=1}^{n}\left(y_i\vec{\theta}^T x_i - e^{\vec{\theta}^T x_i} - \log(y_i!)\right)}$$

$$\nabla LL(\vec{\theta}|\vec{xy}) \overset{set}{=} \text{zero} \rightarrow \frac{\partial LL(\vec{\theta}|\vec{xy})}{\partial\vec{\theta}} = \sum_i y_i x_i - x_i e^{\vec{\theta}^T x_i}$$ (no closed form soln.)

(b) Assuming a MLE $\hat{\theta}$ is solved, predicting a y given x^* follows the form:

The probability of a y given x^* w/ parameters $\hat{\theta}$

Predicted $\eta = \log\lambda$

$\lambda = \theta^T x$

$$p(y|x^*;\hat{\theta}) = \frac{\lambda^y}{y!}e^{-\lambda} = \boxed{\frac{e^{y\hat{\theta}^T x^*}\, e^{-e^{\hat{\theta}^T x^*}}}{y!}}$$

To find the most likely y, iterate from $y=0$ to $n-1$ and return the max from that search (or set $\nabla \rightarrow 0$)

(c) X^* contains a dimension not in $\mathbb{R}^p$, returning $x^* \cdot \hat{\theta}^T = 0$.

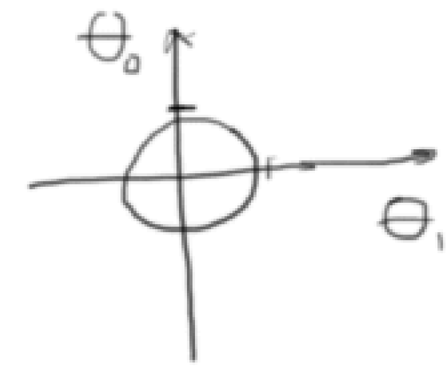
original prediction of $y \Rightarrow \frac{e^{y(0)}e^{(0)}}{y!} = \frac{1}{y!}$, which is unintended.

L2 ridge regression adds $\alpha\|\theta\|^2$

← to differentiate from Poisson?

If we are predicting a maximum likelihood $\hat{\theta}$ first (with $\vec{xy}'$ being $\vec{xy}$ with x^*)

$$\text{loss func } \mathcal{L}_{L2}(\theta|\vec{xy}') = \sum_{i=1}^{n}\left(y_i\vec{\theta}^T x_i - e^{\vec{\theta}^T x_i} - \log(y^!)\right) + \alpha\|\theta\|^2$$

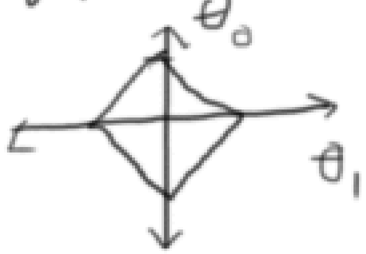


$$\nabla\mathcal{L} = \frac{\partial\mathcal{L}}{\partial\theta} = \sum_{i=1}^{n}\left(\left(y_i x_i - x_i e^{\vec{\theta}^T x_i} - 0\right)\right) + 2\alpha\theta = \begin{bmatrix}\frac{\partial\mathcal{L}}{\partial\theta_1}\\ \vdots \\ \frac{\partial\mathcal{L}}{\partial\theta_p}\end{bmatrix}$$

For $x_i = x^*$, $\mathcal{L}_{x^*} = y_i x^* - x^* e^{(0)} = y_i x^* - x^*$, also unintended.

Additionally, there are not enough data points with the orthogonal dimension to estimate a sensible weight in $\vec{\theta}$, but the weight will approach to Ø, never to Ø itself, and affect all other predictions.

(d) Since we seek to bring the weight of an irrelevant/untenable θ_j to absolute Ø L1 lasso regression offers this by penalizing parameters that do not significantly help the loss function. It trades off inclusion for relevance, by requiring a selection of parameters



Problem 3 – Very Random Forest

Given;
- $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
- $x_i \in \mathbb{R}^p$
- $y_i \in \mathbb{R}$
- sample one feature
- sample $m < n$ x_j's $(\tilde{x})$

(a) Find: Data distributions where this "very" random forest has no bias.

Bias is expected difference between:
- $\mathbb{E}(y|x)$
- $\mathbb{E}_{\mathcal{D}\sim p}[f(x), \mathcal{D}]$

squared.

∴ the single feature trees can account for functions with terms that involve a single feature each in the true distribution, because the ensemble averages the single feature trees together. Assuming ∞ trees.

e.g. $f(\vec{x}) = x_1 + x_2 + \ldots + x_n$
or $f(\vec{x}) = x_1^{k_1} + x_2^{k_2} + \ldots + x_n^{k_n}$ where $k_i \in \mathbb{R}$

(b) With an infinite amount of trees, results in both regular RF and vRF can theoretically reach the truth of (a)'s distribution.

x_i

avg $f(x)_i = x_i^k$

	BIAS	VARIANCE
RF	none	higher
vRF	none	lower

Further discussion on bias and variance.

$$\text{variance} \equiv \mathbb{E}\left[(\text{long term-performance} - \text{expected test performance})^2\right]$$

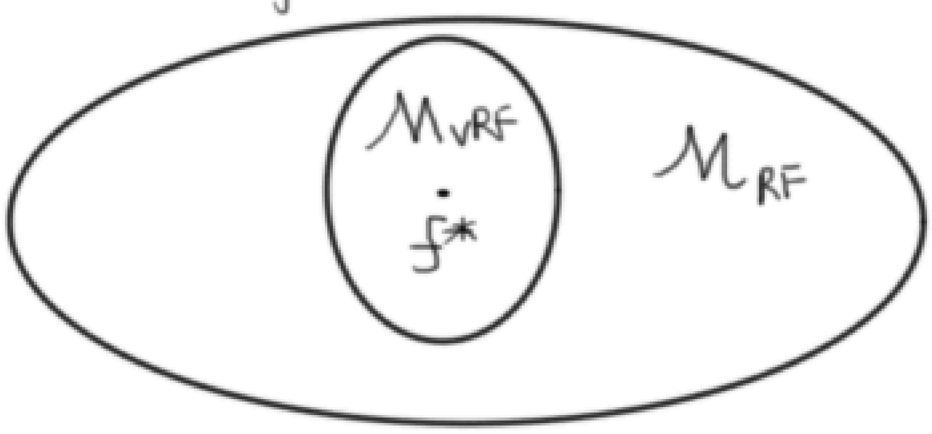
$$= \mathbb{E}_{\mathcal{D}\sim p}\left[\left(\mathbb{E}_{\mathcal{D}\sim p}[f(x);\mathcal{D}] - f(x;\mathcal{D})\right)^2\right]$$

With an infinite amount of trees, the long term performance should closely capture the true distribution described. Test performance ensembled from single feature trees will be accurate, as trees → inf, of course if data captures (will be noise). However, traditional RF trees are less stable in terms of this distribution, since they can introduce small interfeature dependencies which do not actually exist. But with trees approaching infinity, variance too should approach zero (again, only in distributions where vRF has Ø bias.)

We can say vRF approaches Ø variance faster than RF.

In both cases, forest will overfit if not given new data, thus leading to variance,

In a diagram:



$f^* \equiv$ true function

$M_{vRF} \equiv$ model of very Random Forest ∞

$M_{RF} \equiv$ model of traditional Random Forest ∞

M_{RF} more robust but can have more overall variance

M_{vRF} = less robust and more likely to have small variance given that answer is in model space.

Trivially, uniform distributions also have Ø bias, and variance for both models, given $x^* \in$ domain.

Problem 4 – Alternative Losses

Given:
- $\vec{xy}$ $\{(x_1,y_1)\ldots(x_n,y_n)\}$
- $x_i \in \mathbb{R}$ • $y_i \in \{0,1\}, i=1\ldots n$
- $P(y_i=1|x_i) = \sigma(\beta x_i) = 1/(1+e^{-\beta x_i})$
- $P(y_i=0|x_i) = 1-\sigma(\beta x_i)$

(a) Find derivative of squared loss wrt β at $\beta=3$

$x_{n+1}=100$
$y_{n+1}=0$

$$\frac{\partial \sigma(\beta x_{n+1})}{\partial \beta} = \left[y_{n+1} - \sigma(\beta x_{n+1})^2\right]'$$

$$= 2\left(y_{n+1} - \frac{1}{1+e^{-\beta x_{n+1}}}\right)\frac{\partial}{\partial \beta}\left[y_{n+1} - \frac{1}{1+e^{-\beta x_{n+1}}}\right]$$

$$= \left(2x_{n+1}e^{\beta x_{n+1}}\left((y_{n+1}-1)e^{\beta x_{n+1}} + y_{n+1}\right)\right)/(e^{\beta x_{n+1}}+1)^3$$

$$= \frac{2(100)e^{3(100)}(-e^{300})}{(e^{300}+1)^3} = 1.029\times10^{-128} \text{ for } P(y=1)$$

negative for $P(y=0)$

(b) $L(\beta) = \prod_{i=1}^{n} p(Y=y_i \mid x=x_i) = \prod_{i=1}^{n} \sigma(\beta)^{y_i} \cdot (1-\sigma(\beta x))^{(1-y_i)}$

$\therefore \underset{\text{log likelihood}}{\ell(\beta)} = \sum_{i=1}^{n} y_i \log \sigma(\beta x_i) + (1-y_i)\log(1-\sigma(\beta x_i))$

$\underset{\text{loss}}{\mathcal{L}} \equiv -\ell(\beta) \quad \therefore \quad \frac{\partial \mathcal{L}(\beta)}{\partial \beta_j} \underset{\text{chain rule}}{\Rightarrow}$ where $\sigma_{\beta x} \equiv \sigma(\beta \vec{x})$ $\quad \frac{\partial \mathcal{L}(\beta)}{\partial \sigma_{\beta x}} \cdot \frac{\partial \sigma_{\beta x}}{\partial \beta_j}$

$\underset{\text{chain rule}}{\Rightarrow}$ where $\gamma \equiv \beta x$ $\quad \frac{\partial \mathcal{L}(\beta)}{\partial \sigma_{\beta x}} \cdot \frac{\partial \sigma_{\beta x}}{\partial \gamma} \cdot \frac{\partial \gamma}{\partial \beta_j}$

echoing the logit function $\log\left(\frac{p}{1-p}\right)$

$\underset{\text{substitution with derived results}}{\Rightarrow} -\left(\frac{y}{\sigma_{\beta x}} - \frac{1-y}{1-\sigma_{\beta x}}\right) \cdot \sigma_{\beta x}(1-\sigma_{\beta x}) \cdot x_j$

$\underset{\text{expanding}}{\Rightarrow} -(y(1-p) - p(1-y)) \cdot x_j$

$\Rightarrow -(y-p)x_j = -(y-\sigma(\beta x))x_j$

$\underset{\substack{x=100 \\ y=0 \\ \beta=3}}{\Rightarrow} -\left(0 - \frac{1}{1+e^{3(100)}}\right)(100) \cong -(0-1)(100) = 100$

(c) These loss functions imply drastically different characteristics of the parameter β. This is because different distributions call for different concepts of normalization. For some distributions, such as Bernoulli, least squares does not make the most sense because all y points will be a set distance from the line (0 or 1). Bernoulli uses logistic regression as seen in this problem. This is why least squares yields an arbitrary ~0 derivative. In this case, $\mathcal{L} \equiv -\ell$ yields a more palpable result. $\therefore$ do not take $\mathcal{L}$ for granted!