MLPR so far

Train val test splits, generalization, Gaussian stats, CLT

Pre-processing: look at data, one hot encoding, taking logs

Bayes classifiers + Gaussians

Linear regression
- with basis fn's
- regularization
- Bayesian prediction
- ∞ basis fn's: GPs

"Generative model"

Today: revisit "Discriminative models"
fit \( p(y|x) \) directly

Start of non-linear models, fitted with gradient methods

Later will do Bayesian prediction with non-linear models

2019 16 1
Regressing on Labels

\[ f(x) \approx p(y=1|x) \]

(If enough data and basis functions, linear least squares works!)

Often bad idea:

Least squares quadratic fit

Function can give good labels

Terrible square error
Gradients for least squares cost

Residuals
\[ \mathbf{r} = \mathbf{y} - \mathbf{Xw} \]

\[ \mathbf{f} = \mathbf{r} \]

Cost
\[ \mathbf{r}^T \mathbf{r} = (\mathbf{y} - \mathbf{Xw})^T (\mathbf{y} - \mathbf{Xw}) \]

\[ = \mathbf{y}^T \mathbf{y} - (\mathbf{Xw})^T \mathbf{y} - \mathbf{y}^T (\mathbf{Xw}) + \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} \]

"Gradient" vector of partial derivatives
\[ \nabla \mathbf{w} [\mathbf{r}^T \mathbf{r}] = \mathbf{0} - 2 \mathbf{X}^T \mathbf{y} + 2 \mathbf{X}^T \mathbf{X} \mathbf{w} \]

\[ = -2 \mathbf{X}^T (\mathbf{y} - \mathbf{Xw}) \]

Gradient Descent

Initialise \( \mathbf{w} \) (to 0?)

for \( t = 1 \ldots T \):

\[ \mathbf{w} \leftarrow \mathbf{w} - \eta \nabla \mathbf{w} [\mathbf{r}^T \mathbf{r}] \]

\[ \eta \] "eta", small number 0.01?
(Scratch working)

\[
\nabla_w [w^T h] = \begin{bmatrix}
\frac{\partial w^T h}{\partial w_1} \\
\frac{\partial w^T h}{\partial w_2} \\
\vdots \\
\frac{\partial w^T h}{\partial w_0}
\end{bmatrix} = h
\]

some vector

\[
\frac{\partial w^T h}{\partial w_i} = \frac{\partial}{\partial w_i} \sum_j w_j h_j = \frac{\partial}{\partial w_i} (w_i h_i + w_2 h_2 + \ldots + w_i h_i + \ldots + w_0 h_0) = h_i
\]

Matrix Cookbook

2019 L16 ∙
Normal Equations approach

\[ W \left[ I^{T} I \right] = 0 \] at least squares solution

If all weights are happy where they are

\[
\begin{align*}
(X^T X)w &= X^T y \\
w &= (X^T X)^{-1} X^T y
\end{align*}
\]

\[ X^{-1} y \quad \text{Pseudo Inverse} \]

\[ X^{-1} y \quad \text{Inverse} \]

\[ X \quad \text{NxD} \]

"X \backslash y"
When is there a unique solution?

We need $N \geq D$ for unique solution.

One-hot encoding in $\mathbb{R}$

- "red" → 100
- "blue" → 010
- "green" → 001

$x_1, x_2, x_3, x_0$

Imagine I have weights $w$

New weights $\tilde{w}$:

- $\tilde{w}_1 = w_1 + \delta$
- $\tilde{w}_2 = w_2 + \delta$
- $\tilde{w}_3 = w_3 + \delta$
- $\tilde{w}_0 = w_0 - \delta$

Same function

$\tilde{w}^T x = \tilde{w}^T x$ for all $x$

(so same train cost)

$x_1 + x_2 + x_3 + x_0 = 2$

$x_3 = 2 - x_1 - x_2 - x_0$
Logistic Regression
\[ f(x; w) = \sigma(w^T x), \quad f \in [0,1] \]
\[ = \frac{1}{1 + e^{-w^T x}} \]
\[ \sigma(a) = \frac{1}{1 + e^{-a}} \]

Loss Function

Could use square loss again:
\[ \sum_{n=1}^{N} (y^{(n)} - f(x^{(n)}; w))^2 \]

Interpretation:
\[ P(y = 1 \mid x) = f(x; w) \]

Maximum Likelihood, maximize prob. of the data given \( w \):
\[ P(y \mid X, w) = \prod_{n} p(y^{(n)} \mid x^{(n)}, w) \]

Or minimize negative log probability:
\[ \text{NLL} = - \sum_{n: y^{(n)} = 1} \log \sigma(w^T x) - \sum_{n: y^{(n)} = 0} \log(1 - \sigma(w^T x)) \]
I like to make labels $\xi = -1, +1$

$$\xi^{(n)} = 2y^{(n)} - 1$$

Useful fact:

$$1 - \sigma(a) = \sigma(-a)$$

$$\text{NLL} = -\sum_{n=1}^{N} \log \sigma\left(\mathbf{\xi}^{(n)} \mathbf{w}^{(m)} \mathbf{X}^{(m)}\right)$$

Probability of being correct, $\eta_n$

$$\nabla_{\mathbf{w}} \text{NLL} = -\sum_{n=1}^{N} \nabla_{\mathbf{w}} \log \eta_n$$

$$= -\sum_{n=1}^{N} \frac{1}{\eta_n} \nabla_{\mathbf{w}} \eta_n$$

(Chain rule)

$$\frac{\partial \sigma(a)}{\partial a} = \sigma(a) (1 - \sigma(a))$$

$$= -\sum_{n=1}^{N} \frac{1}{\eta_n} \eta_n (1 - \eta_n) \nabla_{\mathbf{w}} \mathbf{z}^{(n)} \mathbf{w}^{(m)} \mathbf{X}^{(m)} \mathbf{X}^{(m)}$$

$$= -\sum_{n=1}^{N} (1 - \eta_n) \mathbf{z}^{(n)} \mathbf{X}^{(n)}$$