today: linear regression

linear regression: closure/motivate with conditional approach to regression \((Y \in \mathbb{R})\)

\[
p(y \mid x; w) = N(y \mid <w, x>, \sigma^2)
\]

\[
x \in \mathbb{R}^d
\]

\[
w \in \mathbb{R}^d
\]

\[
equivalently: \quad Y_i = w^T X_i + \varepsilon_i \quad \text{where} \quad \varepsilon_i | X_i \sim N(0, \sigma^2)
\]

aside: we use "offset" notation for \(x\)

\[
x = \left( \begin{array}{c} x \\ 1 \end{array} \right), \quad \tilde{x} \in \mathbb{R}^{d+1}, \quad \text{"constant feature"}
\]

\[
\tilde{x} \in \mathbb{R}^{d+1}, \quad \mathbb{E} \tilde{x} = w^T \tilde{x} + \varepsilon
\]

\[
\text{"bias"/"offset"
\]

\[
\text{offset (}\tilde{x}_1, \varepsilon\text{)}_{i=1}^n \quad X_i \sim \text{whatever (don't care?)}
\]

\[
y_i | x_i \sim N(w^T x_i, \sigma^2)
\]

conditional likelihood

\[
p(y \mid x; w) = \prod_{i=1}^n p(y_i \mid x_i; w)
\]

\[
\log \left( \prod_{i=1}^n p(y_i \mid x_i; w) \right) = \sum_{i=1}^n \left[ -\frac{(y_i - w^T x_i)^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2) \right]
\]

\[
\frac{\partial}{\partial (\sigma^2)} \left( \right) = 0 \Rightarrow \sum_{i=1}^n \left[ -\frac{(y_i - w^T x_i)^2}{2\sigma^2} \right] = 0
\]

\[
\Rightarrow \frac{\sum_{i=1}^n (y_i - w^T x_i)^2}{\sigma^2} = \frac{n}{\sigma^2} \left( y_i - \bar{w}^T \bar{x} \right)^2
\]

\[
\lim_{\sigma \to 0} \text{obj} \to \infty
\]

\[
\text{obj} \to -\infty \text{ as } \sigma \to 0 \quad \text{so conclude that} \quad \text{this is correct}
\]

\[
\text{global max in } \sigma^2 \text{ for } w \text{ fixed}
\]

(see also)
"design matrix" $\mathbf{X} = \begin{pmatrix} x_1 \cdots x_n \end{pmatrix}$  \( n \times d \) matrix

vector $\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$

\[ Xw = \begin{pmatrix} x_1^T w \\ x_2^T w \\ \vdots \\ x_n^T w \end{pmatrix} \in \mathbb{R}^{n \times 1} \]

\[
\frac{1}{n} \sum_{i=1}^{n} (y_i - w^T x_i)^2 = \| y - \mathbf{X}w \|_2^2
\]

can rewrite $-\log p(y | n, \mathbf{X}) = \| y - \mathbf{X}w \|_2^2 + \text{const}$

\[
\text{MC}L = \min_{w} \| y - \mathbf{X}w \|_2^2 \iff \text{proj. } y \text{ on the column space of } \mathbf{X}
\]

\[
\hat{w}_{\text{MLE}} = \arg \min_{w} \| y - \mathbf{X}w \|_2^2 \quad \text{"least square"}
\]

\[
\begin{align*}
\text{algebra:} & \quad \text{want } \nabla_w \frac{1}{n} \sum_{i=1}^{n} (y_i - w^T x_i)^2 = 0 \\
& \frac{1}{n} \sum_{i=1}^{n} (y_i - w^T x_i)(y_i - w^T x_i) = 0 \\
& \frac{1}{n} \sum_{i=1}^{n} y_i^2 - 2y_i^T X w + w^T X^T X w = 0 \\
& 0 - 2x_i^T y + 2x_i^T x w = 0 \\
& \Rightarrow (X^T X) w = X^T y
\end{align*}
\]

"normal equation"

\[
\hat{w}_{\text{MLE}} = (X^T X)^{-1} X^T y
\]

\[
\text{prediction on training set: } \hat{y} = X \hat{w} = \left( X(X^T X)^{-1} X^T \right) y
\]

\[
\text{for } w \text{ fixed, } \text{rank}(X) \leq \min(n,d)
\]

\[
\text{rank}(X^T X) = \text{rank}(X) \leq \min(n,d)
\]

\[
X^T X \text{ is invertible } \Rightarrow \min(n,d) = d
\]

\[
\Rightarrow \| w \|_2 = d
\]
\begin{align*}
\begin{array}{c}
\text{projection matrix on} \\
\text{column space of } \mathbf{X} \\
\text{(recall geometric projection)}
\end{array}
\end{align*}

\text{Q: if } n < d \text{ (i.e. high dimension) then } \mathbf{X}^\top \mathbf{X} \text{ is not invertible}

\text{or low data regime}

\text{b) if } \mathbf{X}^\top \mathbf{X} \text{ is not invertible } \rightarrow \text{there is no unique } \mathbf{x}^+ \text{ for}

\text{any } \mathbf{w} \text{ s.t. } (\mathbf{X}^\top \mathbf{X}) \mathbf{\hat{w}} = \mathbf{X}^\top \mathbf{y} \text{ is a NCL estimate}

\text{Could choose } \mathbf{\hat{w}} = \text{argmin } \| \mathbf{w} \|_2 \quad \Rightarrow X^+ y \quad \text{pseudo-inverse}

(\text{see Wikipedia})

\begin{align*}
\mathbf{X} &= U \Sigma V^T \\
\mathbf{X}^\top &= V \Sigma^+ U^T
\end{align*}

\text{X}^\top = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X} \text{ when}

\begin{cases}
\mathbf{X} \text{ is full rank} \\
\text{otherwise}
\end{cases}

\begin{align*}
\sum_{i=1}^n \begin{pmatrix} 0 & 0 \\
0 & \sigma_i^2
\end{pmatrix} & \leq \sum_{i=1}^n \begin{pmatrix} 0 & 0 \\
0 & \sigma_i^2
\end{pmatrix} \leq \sum_{i=1}^n \begin{pmatrix} o_i^+ & 0 \\
0 & o_i^+
\end{pmatrix} o_i^+ o_i \Rightarrow \text{not stable}
\end{align*}

\text{Problem: pseudo-inverse is not numerically stable}

\text{instead it is better to regularize to get similar effect}

\text{16:100}

\text{Regularization: (can be motivated from MAP point of view)}

\text{Suppose we put a prior } p(\mathbf{w}) = N(\mathbf{w} | \mathbf{0}, \sigma_w^2 \mathbf{I})

\text{\text{regularizer} } \Rightarrow \text{Laplace prior}

\text{log posterior: } \text{log } p(\mathbf{w} | \text{data}) = \text{log } p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) + \text{log } p(\mathbf{w}) + \text{Cst.}

= -\frac{1}{2\sigma_w^2} \| \mathbf{y} - \mathbf{X} \mathbf{w} \|^2 + \frac{\lambda}{2} \| \mathbf{w} \|^2 + \text{Cst.}

\text{MAP has }
\mathbf{\hat{w}}_{\text{MAP}} = \text{argmin } \| \mathbf{y} - \mathbf{X} \mathbf{w} \|^2 + \frac{\lambda}{2} \| \mathbf{w} \|^2 \Rightarrow \text{ridge regression}

\text{same as "regularized"}

\text{empirical risk minimizer, with squared loss } \text{L}(\mathbf{y} \mid \mathbf{X}, \mathbf{w})

\text{same as } \sum_{i=1}^n \frac{1}{2} (\mathbf{y}_i - \mathbf{w}^\top \mathbf{x}_i)^2

\text{this is } f(\cdot) \text{ is not strongly convex } 
\Rightarrow \text{a unique } \mathbf{w}

\text{f(\cdot) is strongly convex } \Rightarrow \text{a unique } \mathbf{w}
$\nabla w = 0 \Rightarrow (X^TX + \lambda I) w = X^Ty$

Good practice: to either standardize features i.e., make each feature zero mean or normalize them, or make $x_i$ unit norm $\|x_i\|_2 = 1$ or scale features to $[0,1]$ or $[-1,1]$

logistic regression

setup: binary classification $Y \in \{0, 1\} \quad X \in \mathbb{R}^d$

generative model motivation:

Suppose only assumption is there exists a pdf (density) in $\mathbb{R}^d$

for each class conditionals

$p(Y=1 | X=x) = \frac{p(Y=1, X=x)}{p(Y=1, X=x) + p(Y=0, X=x)}$ 

$p(Y=1, X=x) = \log \frac{p(Y=1) p(x)}{p(Y=0) p(x)}$ where $f(x) = \log \frac{p(Y=1) p(x)}{p(Y=0) p(x)}$

prior odds ratio

class-conditional ratio

In general,

$p(Y=1 | X=x) = \sigma(f(x))$

where $\sigma(z) = \frac{1}{1 + \exp(-z)}$ sigmoid function

some properties of $\sigma(z)$:

$\sigma(-z) = 1 - \sigma(z) \quad [\sigma(z) + \sigma(-z) = 1]$

$\frac{d}{dz} \sigma(z) = \sigma(z)(1-\sigma(z))$

To motivate binary logistic regression, consider class conditionals
\[
p(x | m) \triangleq h(x) \exp(\mathbf{m}^T \mathbf{x} - A(m))
\]

Gaussian: \[\log p(x | \mu, \sigma^2) = -\frac{1}{2} \log \sigma^2 - \frac{(x - \mu)^2}{2\sigma^2}\]
\[
\frac{1}{\sqrt{2\pi}\sigma^2} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)
\]

Let \[\mathbf{T}(x) = \begin{bmatrix} -x^2 \\ x \end{bmatrix}\]
\[\mathbf{m}(\mu, \sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2} \\ \mu \end{bmatrix}\]
\[A(m) = \frac{1}{2} \log(\sigma^2) + \frac{\mu^2}{\sigma^2}\]

\[p(x | y=1) = p(x | m_1)\]
\[p(x | y=0) = p(x | m_0)\]

\[\log \text{odds} \quad f(x) = \log \frac{p(x | y=1)}{p(x | y=0)} = \log \frac{p(x | m_1)}{p(x | m_0)} + \frac{1 - \pi}{\pi}
\]
\[= (m_1 - m_0)^T \mathbf{T}(x) + A(m_0) - A(m_1) + \log \frac{1}{1-\pi}
\]
\[= \mathbf{W}^T \phi(x)
\]

where \[\mathbf{W} = \begin{pmatrix} m_1 - m_0 \\ (A(m_0) - A(m_1) + \log \frac{1}{1-\pi}) \end{pmatrix}\]
\[\phi(x) = \begin{pmatrix} \mathbf{T}(x) \\ 1 \end{pmatrix}\]

GOT logistic regression model \[p_w(y=1 | x=x) = \sigma \left( \mathbf{W}^T \phi(x) \right)\]

decision boundary \[1 \iff \mathbf{W}^T \phi(x) > 0\]

exercise to read: try argument above with \(p(x | y) = \mathbf{N}(x | \mu_y, \Sigma_y)\)

- if \(x_0 = x_0\), then \(\phi(x) = \begin{pmatrix} x \\ 1 \end{pmatrix}\)
- otherwise \(\phi(x) = \begin{pmatrix} x \times x^T \end{pmatrix}\)
- note about \( \sigma^2 \) being a global max

(aside: showing that the \( \sigma^2 \) above is the global max is subtle because the objective is not concave in \( \sigma^2 \). I give more info here for your curiosity, but it is not required for the assignment.)

- Formally, to find a global max of a differentiable objective, you need to check all stationary points (zero gradient points), as well as the values at the boundary of the domain. Thus here, you would need to show that the objective cannot take higher value anywhere at the boundary of the domain (which is the case here (exercise!), as the objective goes to \(-\infty\) at the boundary), so you are done (this is the only possible global optimum -- a maximum here, as it should be, given that there are no other stationary points and all values are lower at the boundary, but one could also explicitly check the Hessian to see that it is strictly negative definite at the stationary point, i.e. it looks like a local maximum).

Note that we will see later in the class that the Gaussian is in the exponential family, with a log-concave likelihood in the right ("natural") parameterization, and thus using the invariance principle of the MLE, we could also easily deduce the MLE in the "moment" parameterization which is the usual \((\mu, \sigma^2)\) one, without having to worry about local optima...

- for a cute counter-example illustrating that a differentiable function could have only one stationary point which is a local min but *not a global min* (and thus why one need to look at the values at the boundary), see:
  - https://en.wikipedia.org/wiki/Maxima_and_minima#Functions_of_more_than_one_variable
  - i.e. \( f(x, y) = x^2 + y^2(1 - y)^2 \), \( x, y \in \mathbb{R} \), shows its only critical point is at \((0,0)\), which is a local minimum with \( f(0,0) = 0 \). However, it cannot be a global one, because \( f(2,3) = -5 \).
  - (see picture of function [here](https://en.wikipedia.org/wiki/Maxima_and_minima#Functions_of_more_than_one_variable)
  - (and note that the "Mountain pass theorem" which basically says that if you have a strict local optimum with another point somewhere with the same value, then there must be a saddle point somewhere (a "mountain pass") i.e. another stationary point, does *not hold for this counter-example as one of the required regularity condition, the "Palais-Smale compactness condition" fails. Here, the saddle point (which should intuitively exist) "happens at infinity", which is why it only has one stationary point despite \((0,0)\) not being a global minimum)
  - the moral of the story: intuitions for multivariate optimization are often misleading! (this counter-example would not work in 1d because of Rolle's theorem)