- Parametrized ML models (GLMs)
- Principle of maximum likelihood estimation (MLE)
- Decomposing risk of ML models into:
  - Approximal error, generalization error, optimization error
Last class

\[ p(y|x) \] - captures everything about the dependence

\( b/w \) \( y \& x \)

\[ \mathbb{E}[y|x] \] - Bayes optimal (in least squares loss)

point estimate of \( y \) given \( x \)

Issues: either of these could be arbitrarily complicated.

To deal with this, we use \underline{parametrized models},

\[ p(y|x) = f_\Theta(x) \]

where \( \Theta \in \mathbb{R}^p \) is a

\underline{parameter vector}

\[ = f(x; \Theta) \]
Ex: OLS

\[ y = \beta^T x + \varepsilon \quad \text{where } \varepsilon \sim N(0, \frac{1}{n}) \]

so \[ y | x \sim N(\beta^T x, \frac{1}{n}) \]

or \[ p(y | x) = \frac{1}{\sqrt{2\pi} \sigma} \exp \left( -\frac{(y - \beta^T x)^2}{2\sigma^2} \right) \]

where \( \sigma = \frac{1}{\sqrt{n}} \)

\[ \mathbb{E}[y | x] = \beta^T x \] is our regression function and we estimate \( \hat{\beta} = X^+ y \) from our training data

\[ X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \]

and showed last time that \[ \mathbb{E}[\| \hat{\beta} - \beta \|^2_2 | X] = O \left( \frac{d}{n} \right) \]
OLS is appropriate when data looks like, e.g.

\[ y \]
\[ x \]

\[ y = \begin{cases} 1 & \text{if } x = 0 \\ -1 & \text{if } x = 1 \end{cases} \]

not appropriate for OLS when \( y \in \{-1, 1\} \)
What about models for different types of data?

- meaning, e.g. $y$ is not a continuous r.v.
- $y$ has limited range
- $y$ - $E[y|x]$ is not independent of $x$

Exs.
- Poisson/"shot noise" model

 corresponds to $\text{Poisson}(\lambda)$ is a r.v. that takes on values in $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$ with pmf given by

$$p(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

where $\lambda$ is called the rate, $\sum_{k=0}^{\infty} p(k)$ with $\mathbb{E}X = \lambda$
We use this in ML for predicting count data
\[ y(x) \sim \text{Poisson}(\lambda(x)) \]

This means
\[ \mathbb{E}[y|x] = \lambda(x) \]

Notice since \( \lambda(x) \) is a rate, it must satisfy \( \lambda(x) > 0 \).

We impose this by taking
\[ \lambda(x) = e^{\theta^T x} \]

so
\[ \mathbb{E}[y|x] = e^{\theta^T x} \]

(n.b. we always assume the covariates \( x \) include the constant feature \( 1 \), i.e. \( x = [x_1, \ldots, x_D, 1] \))
- Bernoulli "noise" model

\[ g(x) \sim \text{Bern}(p(x)) \]

useful for classification (e.g. spam detection)

\[ \mathbb{E}[g(x)] = p(x) \]

and we require \( p(x) \in (0, 1) \)

\[ p(x) = \sigma(\theta^T x) \]

so

\[ g(x) \sim \text{Bern}(\sigma(\theta^T x)) \]

\[ \mathbb{E}[g(x)] = \sigma(\theta^T x) \]

logistic sigmoid

\[ \sigma(x) = \frac{1}{1 + e^{-x}} \]

Very used for binary classification
Categorical regression (multiclass classification) model

\[ y \mid x \sim \text{Categorical} \left( p_1(x), \ldots, p_k(x) \right) \]

where \( k \) is the number of categories and

\[ p \left( y = e_i = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \mid x \right) = p_i(x) \]

Note that we need \( p_i(x) \geq 0 \) and \( \sum_{i=1}^{k} p_i(x) = 1 \).

We do this by taking

\[ p_i(x) = \frac{e^{\Theta_i^T x}}{Z(x)} \]

where \( Z(x) \) is the "normalizing constant".

\( \Rightarrow \)

\[ Z(x) = \sum_{i=1}^{k} e^{\Theta_i^T x} \]
Introduce some notation:

\[ \Theta = \begin{bmatrix} \Theta_1^T \\ \vdots \\ \Theta_K^T \end{bmatrix} \in \mathbb{R}^{k \times d}, \quad 1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^k \]

and if \( x \in \mathbb{R}^k \) then interpret \( e^x = \begin{bmatrix} e^{x_1} \\ \vdots \\ e^{x_K} \end{bmatrix} \)

Then we have

\[ p_{\Theta}(x) = \begin{bmatrix} p_1(x) \\ \vdots \\ p_K(x) \end{bmatrix} = \begin{bmatrix} e^{\Theta_1^T x} \\ \vdots \\ e^{\Theta_K^T x} \end{bmatrix}, \quad \frac{1}{\gamma_{\Theta}(x)} = \frac{e^{\Theta x}}{1^T e^{\Theta x}} \]

and

\[ y \mid x \sim \text{Categorical} \left( p_{\Theta}(x) \right) \]

\[ \mathbb{E} \left( y \mid x \right) = \begin{bmatrix} p_{\Theta}(x)_1 e_1 \\ \vdots \\ p_{\Theta}(x)_K e_K \end{bmatrix} = \begin{bmatrix} p_{\Theta}(x)_1 \\ \vdots \\ p_{\Theta}(x)_K \end{bmatrix} = p_{\Theta}(x) \]
How to interpret $\Theta$?

$$
\frac{p(y = e_i | x)}{p(y = e_j | x)} = \frac{e^{\Theta_i^T x} / Z_\Theta(x)}{e^{\Theta_j^T x} / Z_\Theta(x)} = \frac{e^{\Theta_i^T x}}{e^{\Theta_j^T x}}
$$

$$
= e^{(\Theta_i - \Theta_j)^T x}
$$
In general we can fit an entire class of models, called Generalized Linear Models (GLMs) by choosing

\[ y \mid x \sim \text{Exp}\left(p(x)\right) \]

where \( \text{Exp} \) is a particular kind of exponential family distribution and \( p(x) \) is a parameter vector.

For a given \( \text{Exp} \)

\[ E[y \mid x] = g^{-1}(\Theta^T x) \quad \text{where } g \text{ is called the link function} \]

(e.g. for the Bernoulli model

\[ E[y \mid x] = \sigma(\Theta^T x) \quad \text{so } g^{-1} = \sigma \Rightarrow g(x) = \ln \left( \frac{x}{1-x} \right) \]
Maximum Likelihood Estimation

We fixed a model that we expect to work well for our problem. How do we estimate the parameters of that model?

\[ y_i | x \sim P_\theta(x) \]

and given \( \mathcal{D}(x_i, y_i) \) for \( i = 1 \)

To estimate the optimal parameters \( \theta \) we use the maximum likelihood principle

\[ \hat{\theta} = \arg \max_{\theta} \prod_{i=1}^{n} P_\theta(y_i|x_i) \]
This process is called maximum likelihood estimation. Notice

\[ \hat{\Theta} = \arg \max_\Theta \prod_{i=1}^n p_\Theta(y_i|x_i) \]

\[ = \arg \max_\Theta \left( \prod_{i=1}^n p_\Theta(y_i|x_i) \right)^{\frac{1}{n}} \]

\[ = \arg \max_\Theta \frac{1}{n} \sum_{i=1}^n \log p_\Theta(y_i|x_i) \]

\[ = \arg \min_\Theta \frac{1}{n} \sum_{i=1}^n -\log p_\Theta(y_i|x_i) \]

[called negative log likelihood of \( \Theta \) (NLL)]
MLE for Gaussian model

\[ y = \beta^T x + \epsilon \quad \text{where} \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \]

or equivalently

\[ p_\theta(y_i | x_i) = \frac{1}{\sqrt{2\pi\sigma}} \exp \left( -\frac{(y_i - \beta^T x_i)^2}{2\sigma^2} \right) \]

use MLE to estimate \( \beta \):

\[
\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} -\log p_\theta(y_i | x_i)
\]

\[
= \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} - \left[ \log \left( \frac{1}{\sqrt{2\pi\sigma}} \right) - \frac{(y_i - \beta^T x_i)^2}{2\sigma^2} \right]
\]

\[
= \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \beta^T x_i)^2}{2\sigma^2} + \log \left( \frac{1}{\sqrt{2\pi\sigma}} \right)
\]
\[ \Rightarrow \quad \hat{\beta} = \arg \min_{\beta} \frac{1}{n \sigma^2} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \]

because \( \sigma \) is a constant

\[ = \arg \min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \]

\[ = \arg \min_{\beta} \| X \beta - y \|_2^2 \]

\[ = X^+ y = \beta_{ols} \]

so MLE for Gaussian model leads to OLS