CS-GY 6923: Lecture 3
Regularization + Bayesian Perspective

NYU Tandon School of Engineering, Akbar Rafiey

Course news

- The first written HW has been posted, due Feb 19.
- 10% extra credit will be given if solutions are typewritten using LaTeX, Markdown, etc.
- All solutions must be written independently.
- Reminder: Lab 02 due Feb 12.

Recap

Model selection:

- Train models $f^{(1)}_{\theta_1}, \ldots, f^{(q)}_{\theta_q}$ independently on training data to find optimal parameters $\theta^*_1, \ldots, \theta^*_q$.
- Check loss $L_{test}(f^{(1)}, \theta^*_1), \ldots, L_{test}(f^{(q)}, \theta^*_q)$ on test data.
- Select model with lowest test lost.

Can be used for arbitrary sets of models. Often used when you are not sure how "complex" your model should be for the data, and want to find a sweet spot between a good fit, and not overfitting.

Model selection example

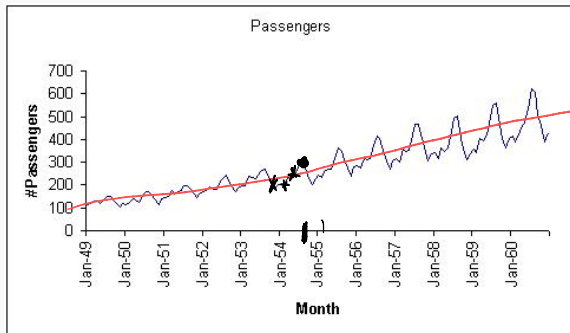**Electrocorticography ECoG (Lab 02):**

- Implant grid of electrodes on surface of the brain to measure electrical activity in different regions.
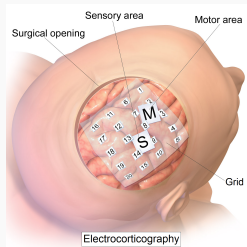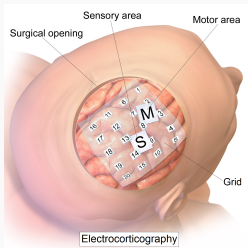

- Predict hand motion based on ECoG measurements.
- **Model order:** predict movement at time t using brain signals at time $t, t - 1, \ldots, t - q$ for varying values of q.

Predicting time t based on a linear function of the signals at time $t, t-1, \ldots, t-q$ is not the same as fitting a line to the time series. It's much more expressive.

Electrocorticography ECoG lab:

First lab where computation actually matters (solving regression problems with $\sim 40k$ examples, ~ 1500 features)

Makes sense to test and debug code using a subset of the data.

> **Slight caveat: The train-test paradigm is typically not how machine learning or scientific discovery works in practice!**

Typical workflow:

- Train a class of models.
- Test.
- Adjust class of models.
- Test.
- Adjust class of models.
- Cont...

Final model implicitly depends on test set because performance on the test set guided how we changed our model.

Adaptive data analysis

Popularity of ML benchmarks and competitions leads to adaptivity at a massive scale.

Kaggle (various competitions)

Imagenet (image classification and categorization)

Is adaptivity a problem? Does it lead to over-fitting? How much? How can we prevent it?

The reusable holdout: Preserving validity in adaptive data analysis

Cynthia Dwork[1,*], Vitaly Feldman[2,*], Moritz Hardt[3,*], Toniann Pitassi[4,*], Omer Reingold[5,*], Aaron Roth[6,*]

+ See all authors and affiliations

Do ImageNet Classifiers Generalize to ImageNet?

Benjamin Recht*   Rebecca Roelofs   Ludwig Schmidt   Vaishaal Shankar
UC Berkeley   UC Berkeley   UC Berkeley   UC Berkeley

12 Jun 2019

Abstract

We build new test sets for the CIFAR-10 and ImageNet datasets. Both benchmarks have been the focus of intense research for almost a decade, raising the danger of overfitting to excessively re-used test sets. By closely following the original dataset creation processes, we test to what extent current classification models generalize to new data. We evaluate a broad range of models and find accuracy drops of $3\% - 15\%$ on CIFAR-10 and $11\% - 14\%$ on ImageNet. However, accuracy gains on the original test sets translate to larger gains on the new test sets. Our results suggest that the accuracy drops are not caused by adaptivity, but by the models' inability to generalize to slightly "harder" images than those found in the original test sets.

9

Imagenet dataset

IM🅰GENET

Explore Download Challenges Publications Updates About

14,197,122 images, 21841 synsets indexed

Not logged in. Login | Signup
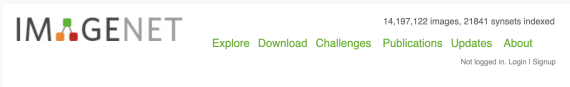
Collected by Fei-Fei Li's group at Stanford in 2006ish and labeled using Amazon Mechanical Turk.

mammal ⟶ placental ⟶ carnivore ⟶ canine ⟶ dog ⟶working dog⟶ husky

vehicle ⟶ craft ⟶ watercraft ⟶ sailing vessel ⟶ sailboat ⟶ trimaran
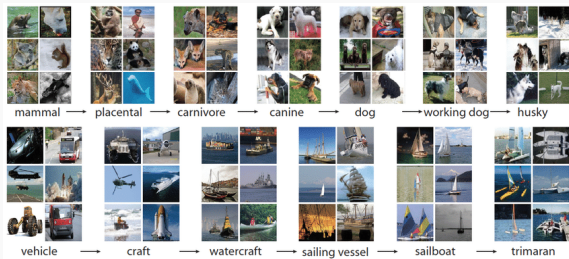
We now have neural network models that can solve these classification problems with $> 95\%$ accuracy.

Do ImageNet Classifiers Generalized to ImageNet?
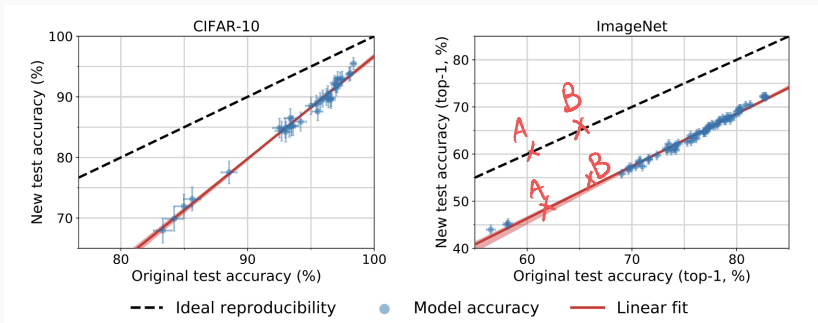
Interestingly, when comparing popular vision models on "fresh" data, while performance dropped across the board, the relative rank of model performance did not change significantly.

Regularization

Over-parameterized models

- In the model selection examples we discussed last class, we had full control over the complexity of the model: could range from underfitting to overfitting.

- In practice, you often don't have this freedom. Even the most basic model might lead to overfitting.

Example: Linear regression model where $d \geq n$.
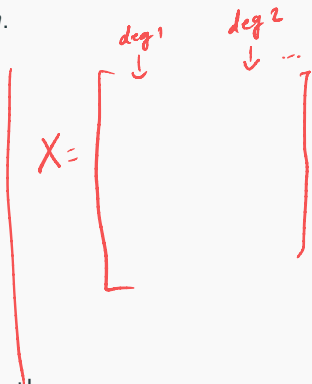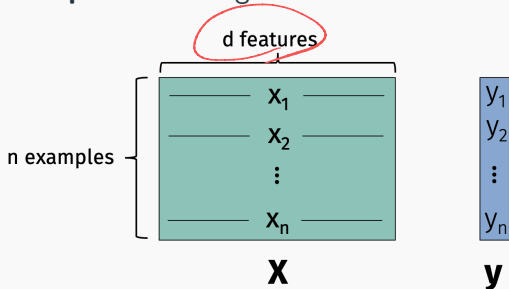
Can (almost) always find $\boldsymbol{\beta}$ so that $\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$ exactly.

Why?

High dimensional linear models

Claim: For almost all sets of n, length n vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$, we can write any vector $\mathbf{y}$ as a linear combination of these vectors.

$$\beta_1 x_{11} + \beta_2 x_{21} + \cdots + \beta_d x_{d1} = y_1$$
$$\beta_1 x_{12} + \beta_2 x_{22} + \cdots + \beta_d x_{d2} = y_2$$
$$\vdots$$

$$X = \begin{bmatrix} \\ \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} \langle x_1, \beta_1 \rangle \\ \langle x_2, \beta_2 \rangle \end{bmatrix}$$

I.e., can find some coefficients so that

$$\beta_1 \mathbf{x}^{(1)} + \ldots + \beta_d \mathbf{x}^{(d)} = \mathbf{X}\beta = \mathbf{y}.$$

$n \times d \quad d \times 1 \quad \rightsquigarrow \quad n \times 1$

Zero train loss

- We will discuss some models later in the class where zero training loss is not necessarily a bad sign: *k*-nearest neighbors, some neural nets.

- Typically however it will be a sign of overfitting, as in the polynomial regression example.

Select some subset of $\ll n$ features to use in model:

$\mathbf{X}$ → $\tilde{\mathbf{X}}$
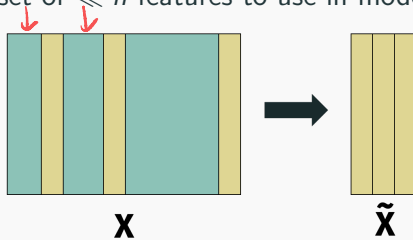
Filter method: Compute some metric for each feature, and select features with highest score.

- Example: compute loss or R^2 value when each feature in $\mathbf{X}$ is used in single variate regression.

Any potential limitations of this approach?

Exhaustive approach: Pick best subset of q features. Train $\binom{d}{q}$ models.

$$O\left(d^{q}\right)$$

$$d^{q} \quad = \quad \frac{d!}{(d-q)! \; q!}$$

$$= \frac{d \cdot (d-1)(d-2)\cdots(d-q)!}{(d-q)! \; q!}$$

Faster approach: Greedily select q features. $O(d \cdot q)$

Stepwise Regression: $d + (d-1) + (d-2) + \cdots + (d-q)$

- **Forward:** Step 1: pick single feature that gives lowest loss. Step k: pick feature that when combined with previous $k - 1$ chosen features gives lowest loss.

- **Backward:** Start with all of the features. Greedily eliminate those which have least impact on model performance.

Feature selection deserves more than two slides, but we won't go into too much more detail!

Alternative approach

Regularization: Discourage overfitting by adding a <u>regularization penalty</u> to the loss minimization problem.

$$\min_{\beta} \left[L(\beta) + Reg(\beta) \right].$$

Regularization: Discourage overfitting by adding a <u>regularization penalty</u> to the loss minimization problem.

$$\min_{\boldsymbol{\beta}} \left[L(\boldsymbol{\beta}) + Reg(\boldsymbol{\beta}) \right].$$

$$\|\beta\|_1 = |\beta_1| + |\beta_2| + \cdots + |\beta_d|$$

Example: Least squares regression: $L(\boldsymbol{\beta}) = \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2$.

$\lambda \geq 0$

- Ridge regression (ℓ_2): $Reg(\boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}\|_2^2$
- LASSO (least absolute shrinkage and selection operator) (ℓ_1): $Reg(\boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}\|_1$
- Elastic net: $Reg(\boldsymbol{\beta}) = \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2$

$\lambda_2, \lambda_1 \geq 0$

Note that $\arg\min_{\boldsymbol{\beta}} \left[L(\boldsymbol{\beta}) + Reg(\boldsymbol{\beta}) \right] \neq \arg\min_{\boldsymbol{\beta}} \left[L(\boldsymbol{\beta}) \right]$

20

Ridge regression: $\min_\beta \left(\|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda\|\beta\|_2^2 \right)$.

- As $\lambda \to \infty$, we expect $\|\beta\|_2^2 \to 0$ and $\|\mathbf{X}\beta - \mathbf{y}\|_2^2 \to \|\mathbf{y}\|_2^2$.
- By choosing different values of λ we have models of varying accuracy/complexity.
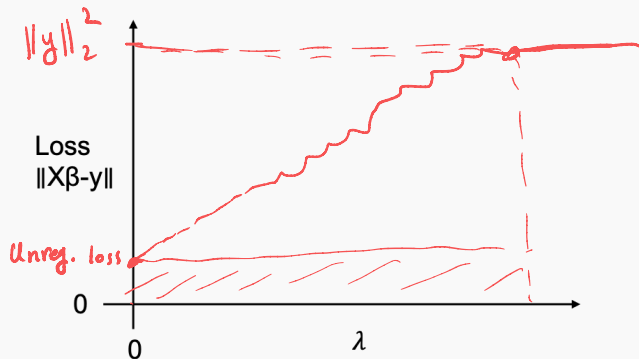
Polynomial examples

Fit degree 20 polynomial with varying levels of regularization.
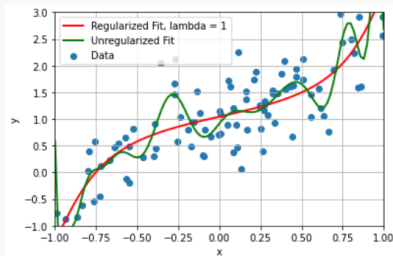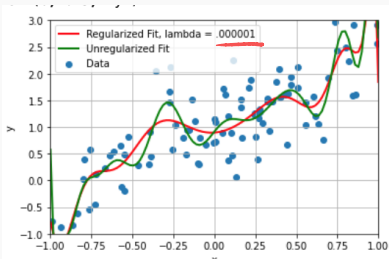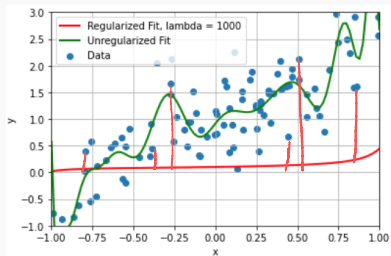
$$\lambda \| \beta \|_2^2$$

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_d \end{bmatrix}$$

$$y_i \approx \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots$$
$$\cdots + \beta_d x_{id}$$

Ridge regression: $\min_{\boldsymbol{\beta}} \left(\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \right).$

- As $\lambda \to \infty$, we expect $\|\boldsymbol{\beta}\|_2^2 \to 0$ and $\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 \to \|\mathbf{y}\|_2^2$.
- Feature selection methods attempt to set many coordinates in $\boldsymbol{\beta}$ to 0. Ridge regularizations encourages coordinates to be close to zero.
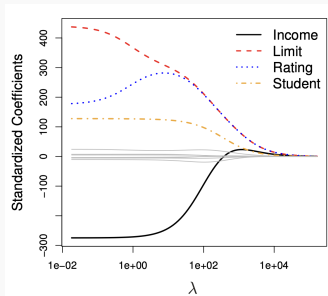
How do we minimize: $L_R(\boldsymbol{\beta}) = \underbrace{\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2}_{} + \underbrace{\lambda\|\boldsymbol{\beta}\|_2^2}_{}$?

$$\nabla L_R(\beta) = 2X^T(X\beta - y) + 2\lambda\beta$$

$$\nabla L_R(\beta) = 0 \Rightarrow 2X^T(X\beta - y) + 2\lambda\beta = 0$$

$$\Rightarrow X^TX\beta - X^Ty + \lambda\beta = 0$$

$$X^TX\beta + \lambda\beta = X^Ty$$

$$\left(X^TX + \lambda \underset{d\times d}{I}\right)\beta = X^Ty$$

$$\Rightarrow \beta = \left(X^TX + \lambda I\right)^{-1}X^Ty$$

$$L(\beta) = \|X\beta - y\|_2^2$$

$$\rightarrow \beta^* = \left(X^TX\right)^{-1}X^Ty$$

24

Lasso regularization

Lasso regularization: $\min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \lambda\|\beta\|_1$.

- As $\lambda \to \infty$, we expect $\|\beta\|_1 \to 0$ and $\|\mathbf{X}\beta - \mathbf{y}\|_2^2 \to \|\mathbf{y}\|_2^2$.
- Typically encourages subset of β_i's to go to zero, in contrast to ridge regularization.
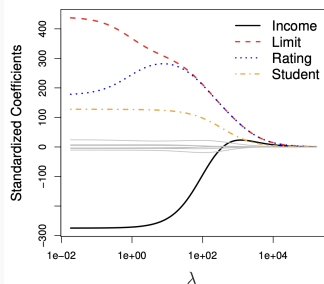
Ridge regularization   Lasso Regularization
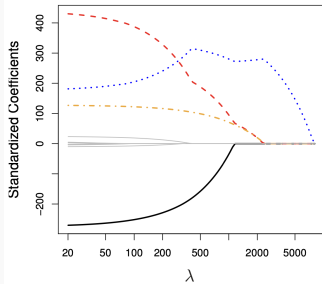
Lasso regularization

Pros:

- Simpler, more interpretable model.
- More intuitive reduction in model order.

Cons:

- No closed form solution because $\|\boldsymbol{\beta}\|_1$ is not differentiable.
- Can be solved with iterative methods, but generally not as quickly as ridge regression.

$$\lambda = 0.1 \quad v.s \quad \lambda = 0.2 \quad \times$$

Notes:

- Model selection/cross validation used to choose optimal scaling λ on $\lambda\|\boldsymbol{\beta}\|_2^2$ or $\lambda\|\boldsymbol{\beta}\|_1$.

- Often grid search for best parameters is performed in "log space". E.g. consider $[\lambda_1, \ldots, \lambda_q] = 1.5^{[-4,-3,-2,-1,-0,1,2,3,4]}$.

- Regularization methods are <u>not invariant</u> to data scaling. Typically when using regularization we mean center and scale columns to have unit variance.

The bayesian/probabilistic modeling perspective

- **Data Examples:** $x_1, \ldots, x_n \in \mathbb{R}^d$
- **Target:** $y_1, \ldots, y_n \in \{0, 1, \ldots, q-1\}$ when there are q classes.
    - Binary Classification: $q = 2$, so each $y_i \in \{0, 1\}$.
    - Multi-class Classification: $q > 2$. [1]

---

[1]Note that there is also multi-label classification where each data example may belong to more than one class.

Classification examples

- Medical diagnosis from MRI: 2 classes.
- MNIST digits: 10 classes.
- Full Optical Character Regonition: 100s of classes.
- ImageNet challenge: 21,000 classes.

  Running example today: Email Spam Classification.

Classification can (and often is) solved using the same **loss-minimization framework** we saw for regression.

We won't see that today! We're going to use classification as a window into another way of thinking about machine learning.

Will give an interesting new justifications for tools like regularization. Will also give us an approach for generative ML.

Rest of Today: ML from a **Probabilistic Modeling/Bayesian Perspective**.

Probabilistic modeling

In a Bayesian or Probabilistic approach to machine learning we always start by conjecturing a

probabilistic model

that plausibly could have generated our data.

- The model guides how we make predictions.
- The model typically has unknown parameters $\vec{\theta}$ and we try to find the most reasonable parameters based on **observed** data (more on this later in lecture).

Typically we try to keep things simple!

✓   ✗   ✗
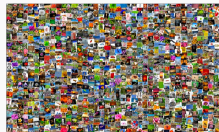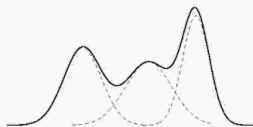
Exercise: Come up with a probabilistic model for the following data set $(x_1, y_1), \ldots, (x_n, y_n)$.

- For n **NYC apartments**: each x_i is the size of the apartment in square feet. Each y_i is the monthly rent in dollars.

What are the unknown parameters of your model. What would be a guess for their values? How would you confirm or refine this guess using data?

$$y_i = c \cdot x_i + \eta$$

$$x_i \quad \text{uniformly from} \quad [300, 1000]$$

Probabilistic modeling

Dataset: $(x_1, y_1), \ldots, (x_n, y_n)$

Description: For n **NYC apartments**: each x_i is the size of the apartment in square feet. Each y_i is the monthly rent in dollars.

Model:

Probabilistic modeling

Dataset: $(x_1, y_1), \ldots, (x_n, y_n)$

Description: For n **students**: each $x_i \in \{$Fresh., Soph., Jun., Sen.$\}$

indicating class year. Each $y_i \in \{0, 1\}$ with zero indicating the
student has not taken machine learning, one indicating they have.

Model:

Naive bayes classifier

Goal:

- Build a probabilistic model for a binary classification problem.

- Estimate parameters of the model.

- From the model derive a classification rule for future predictions (the **Naive Bayes Classifier**).
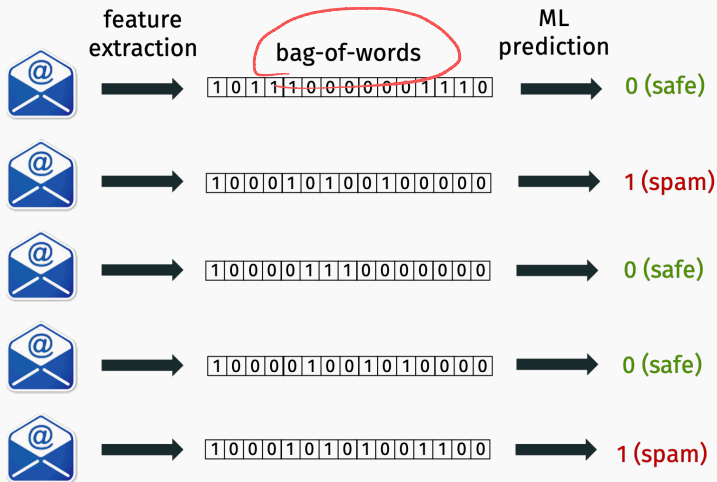
Both target labels and data vectors are binary.

Let's create a model that generates spam and non-spam emails.
Observation: Since bag-of-words features don't care about word order, our model does not need to either.

- Common approach: assign a probability $p_i \in [0, 1]$ to word i. Set $\mathbf{x}_i = 1$ with probability p_i, $\mathbf{x}_i = 0$ with probability $1 - p_i$.

$$p_{\text{the}} = .99 \qquad p_{\text{calendar}} = .07 \qquad p_{\text{toothbrush}} = \frac{17}{11K}$$

38

Email model

How can we make this model richer when we take spam into account?

- Different words tend to be more or less frequent in spam or regular emails.

Not Spam

$$p_{\text{won}} = 0.01$$
$$p_{\$} = 0.02$$
$$p_{student} =$$

Spam

$$p_{\text{won}} = 0.5$$
$$p_{\$} = 0.7$$
$$p_{student} =$$

Probabilistic model for (bag-of-words, label) pair $(\mathbf{x}, y)$:

- Set $y = 0$ with probability p_0, $y = 1$ with probability $p_1 = 1 - p_0$.
  - p_0 is probability an email is not spam (e.g. 99%).
  - p_1 is probability an email is spam (e.g. 1%).
- If $y = 0$, for each i, set $x_i = 1$ with prob. p_{i0}.
- If $y = 1$, for each i, set $x_i = 1$ with prob. p_{i1}.

Unknown model parameters:

- $p_0, p_1,$
- $p_{10}, p_{20}, \ldots p_{d0}$, one for each of the d vocabulary words.
- $p_{11}, p_{21}, \ldots p_{d1}$, one for each of the d vocabulary words.

How would you estimate these parameters?

Parameter estimation

Reasonable way to set parameters:

- Set p_0 and p_1 to the empirical fraction of not spam/spam emails.

- For each word i, set p_{i0} to the empirical probability word i appears in a non-spam email.

- For each word i, set p_{i1} to the empirical probability word i appears in a spam email.

Estimating these parameters from previous data examples is the only "training" we will do.

Done with modeling

on to prediction

- **Probability:** p(x) – the probability event x happens.

- **Joint probability:** p(x,y) – the probability that event x <u>and</u> event y happen.

- **Conditional Probability** $p(x \mid y)$ – the probability x happens <u>given</u> that y happens.

$$P(x,y) = P(y) \cdot P(x|y)$$

$$p(x|y) = \frac{p(x,y)}{p(y)}$$

$$= P(x) \cdot P(y|x)$$

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Proof:

$$p(x,y) = p(y) \cdot p(x|y)$$

$$p(x,y) = p(x) \cdot p(y|x)$$

$$\Rightarrow \quad p(y) \cdot p(x|y) = p(x) \cdot p(y|x)$$

$$\Rightarrow \quad p(x|y) = \frac{p(x) \cdot p(y|x)}{p(y)}$$

Classification rule

Given unlabeled input $(\mathbf{w}, \underline{\quad})$, choose the label $y \in \{0, 1\}$ which is <u>most likely</u> given the data. Recall $\mathbf{w} = [0, 0, 1, \ldots, 1, 0]$.

Classification rule: maximum a posterior (MAP) estimate.

Step 1. Compute:

- $p(y = 0 \mid \mathbf{w})$: prob. $y = 0$ given observed data vector $\mathbf{w}$.
- $p(y = 1 \mid \mathbf{w})$: prob. $y = 1$ given observed data vector $\mathbf{w}$.

Step 2. Output: 0 or 1 depending on which probability is larger.

$p(y = 0 \mid \mathbf{w})$ and $p(y = 1 \mid \mathbf{w})$ are called **posterior** probabilities.

How to compute the posterior? **Bayes rule!**

$$p(y = 0 \mid \mathbf{w}) = \frac{p(\mathbf{w} \mid y = 0)p(y = 0)}{p(\mathbf{w})} \tag{1}$$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \tag{2}$$

- **Prior:** Probability in class 0 <u>prior</u> to seeing any data.
- **Posterior:** Probability in class 0 <u>after</u> seeing the data.

Goal is to determine which is larger:

$$p(y = 0 \mid \mathbf{w}) = \frac{p(\mathbf{w} \mid y = 0)p(y = 0)}{p(\mathbf{w})} \qquad \text{vs.}$$

$$p(y = 1 \mid \mathbf{w}) = \frac{p(\mathbf{w} \mid y = 1)p(y = 1)}{p(\mathbf{w})}$$

- We can ignore the evidence $p(\mathbf{w})$ since it is the same for both sides!

- $p(y = 0)$ and $p(y = 1)$ already known (computed from training data). These are our computed parameters p_0, p_1.

- $p(\mathbf{w} \mid y = 0) = ?$ $p(\mathbf{w} \mid y = 1) = ?$

Consider the example $\mathbf{w} = [0, 1, 1, 0, 0, 0, 1, 0]$.

Recall that, under our model, index i is 1 with probability p_{i0} if we are not spam, and 1 with probability p_{i1} if we are spam .

$$p(\mathbf{w} \mid y = 0) = \quad p(w_0 = 0 \mid y = 0) \cdot p(w_1 = 1 \mid y = 0) \cdots$$

$$(1 - p_{00}) \cdot p_{10} \cdot p_{20} \cdot (1 - p_{30}) \cdots$$

$$p(\mathbf{w} \mid y = 1) = \quad p(w_0 = 0 \mid y = 1) \cdot p(w_1 = 1 \mid y = 1) \cdots$$

$$(1 - p_{01}) \cdot p_{11} \cdot p_{21} \cdots$$

<div align="center">

Final Naive Bayes Classifier

</div>

Training/Modeling: Use existing data to compute:

- $p_0 = p(y = 0), p_1 = p(y = 1)$
- For all i compute:
  - $p_{i0} = p(x_i = 1 \mid y = 0)$ and $(1 - p_{i0}) = p(x_i = 0 \mid y = 0)$
  - $p_{i1} = p(x_i = 1 \mid y = 1)$ and $(1 - p_{i1}) = p(x_i = 0 \mid y = 1)$

Prediction:

- For new input $\mathbf{w}$:
  - Compute $p(\mathbf{w} \mid y = 0) = \prod_i p(w_i \mid y = 0)$
  - Compute $p(\mathbf{w} \mid y = 1) = \prod_i p(w_i \mid y = 1)$
- Return

$$\arg\max \left[p\left(\mathbf{w} \mid y = 0\right) \cdot p\left(y = 0\right), p\left(\mathbf{w} \mid y = 1\right) \cdot p\left(y = 1\right) \right].$$

(handwritten annotation:) if $w_i = 1 \to p_{i0}$
if $w_i = 0 \to 1 - p_{i0}$

Other applications of
the bayesian perspective

The Bayesian view offers an interesting alternative perspective on
<u>many</u> machine learning techniques.

Example: Linear Regression.

Probabilistic model:

$$y = \langle \mathbf{x}, \boldsymbol{\beta} \rangle + \eta$$

where the η drawn from $N(0, \sigma^2)$ is **random Gaussian noise**.
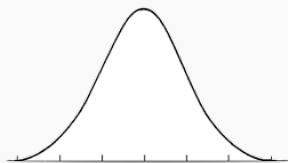
$$Pr(\eta = z) \sim \quad e^{-z^2}$$

The symbol $\sim$ means "is proportional to".

Names for the same thing: Normal distribution, Gaussian distribution, bell curve.

Parameterized by mean μ and variance σ^2.

η is a continuous random variable, so it has a probability density function $p(\eta)$ with $\int_{-\infty}^{\infty} p(\eta)d\eta = 1$

$$p(\eta) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{\eta-\mu}{\sigma})^2} \qquad \sim e^{-\eta^2}$$

51

Gaussian distribution refresher

The important thing to remember is that the the PDF falls off exponentially as we move further from the mean.
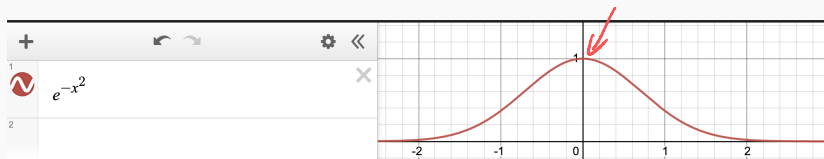
The normalizing constant in front $1/2$, etc. don't matter so much.

$$e^{-0} = 1$$

$$e^{-(2)^2} = \frac{1}{e4}$$

$$e^{-(-2)^2} = \frac{1}{e^4}$$

$$e^{-(8)^2} = \frac{1}{e16}$$

Example: Linear Regression.

Probabilistic model:

$$y = \langle \mathbf{x}, \boldsymbol{\beta} \rangle + \eta$$

where the η drawn from $N(0, \sigma^2)$ is **random Gaussian noise**. The noise is <u>independent</u> for different inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Bayesian regression
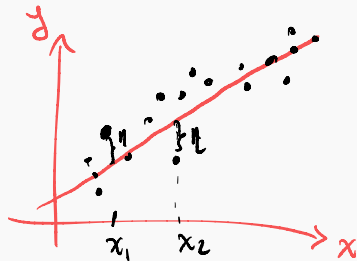
How should we select β for our model?

Also use a Bayesian approach!

Choose β to maximize:

$$\text{posterior} = \Pr(\beta \mid \mathbf{X}, \mathbf{y}) = \frac{\Pr(\mathbf{X}, \mathbf{y} \mid \beta) \Pr(\beta)}{\Pr(\mathbf{X}, \mathbf{y})} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}.$$

In this case, we don't have a prior – no values of β are inherently more likely than others.

Choose β to maximize just the likelihood:

$$\frac{\Pr(\mathbf{X}, \mathbf{y} \mid \beta)\cancel{\Pr(\beta)}}{\cancel{\Pr(\mathbf{X}, \mathbf{y})}} = \frac{\text{likelihood} \times \cancel{\text{prior}}}{\cancel{\text{evidence}}}.$$

This is called the **maximum likelihood estimate**.

Fixed design linear regression

Often we think of **X** as fixed and deterministic, and only **y** is generated at random in the model. This is called the fixed design setting. Can also consider a randomized design setting, but it is slightly more complicated.

In the fixed design setting our task of maximizing $\Pr(\mathbf{X}, \mathbf{y} \mid \boldsymbol{\beta})$ simplifies to maximizing

$$\max_{\boldsymbol{\beta}} \Pr(\mathbf{y} \mid \boldsymbol{\beta})$$

Maximum likelihood estimate

Data:

$$y_i = \langle x_i, \beta \rangle + \eta_i \implies y_i - \langle x_i, \beta \rangle = \eta_i$$

$$\mathbf{X} = \begin{bmatrix} - & \mathbf{x}_1 & - \\ - & \mathbf{x}_2 & - \\ & \vdots & \\ - & \mathbf{x}_n & - \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_d \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Model: $y_i = \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle + \eta_i$ where $p(\eta_i = z) \sim e^{-z^2/2\sigma^2}$ and $\eta_1, \ldots, \eta_n$ are independent.

$$\Pr(\mathbf{y} \mid \boldsymbol{\beta}) \sim \prod_{i=1}^{n} e^{-(y_i - \langle x_i, \beta \rangle)^2/2\sigma^2}$$

$$\Pr(y_i \mid \beta) \sim e^{-(y_i - \langle x_i, \beta \rangle)^2/2\sigma^2}$$

Log likelihood

Easier to work with the **log likelihood**:

$$
\begin{aligned}
\arg\max_{\boldsymbol{\beta}} \Pr(\mathbf{X}, \mathbf{y} \mid \boldsymbol{\beta}) &= \arg\max_{\boldsymbol{\beta}} \prod_{i=1}^{n} e^{-(y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)^2 / 2\sigma^2} \\
&= \arg\max_{\boldsymbol{\beta}} \ \log\left(\prod_{i=1}^{n} e^{-(y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)^2 / 2\sigma^2} \right) \\
&= \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} -(y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)^2 / 2\sigma^2 \\
&= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)^2.
\end{aligned}
$$

Conclusion: Choose $\boldsymbol{\beta}$ to minimize:

$$\sum_{i=1}^{n}(y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2.$$

This is a completely different justification for squared loss!

Minimizing the ℓ_2 loss is optimal in a certain sense when you assume your data follows a linear model with i.i.d. Gaussian noise.

If we had modeled our noise η as Laplace noise, we would have found that minimizing $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_1$ was optimal.
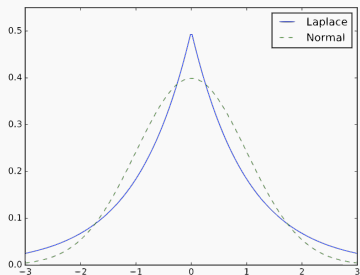
$$Pr(\eta = z) \sim$$

Laplace noise has "heavier tails", meaning that it results in more outliers.

This is a completely different justification for ℓ_1 loss.

Bayesian regularization

We can add another layer of probabilistic modeling by also assuming $\boldsymbol{\beta}$ is random and comes from some distribution, which encodes our <u>prior</u> belief on what the parameters are.

Maximum a posteriori (MAP estimation):

$$\Pr(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}) = \frac{\Pr(\mathbf{X}, \mathbf{y} \mid \boldsymbol{\beta}) \Pr(\boldsymbol{\beta})}{\Pr(\mathbf{X}, \mathbf{y})}.$$

Assume values in $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_d]$ come from some distribution.

- **Common model:** Each β_i drawn from $N(0, \gamma^2)$, i.e. normally distributed, independent.
- Encodes a belief that we are unlikely to see models with very large coefficients.

Bayesian regularization

Goal: choose β to maximize:

$$\Pr(\beta \mid \mathbf{X}, \mathbf{y}) = \frac{\Pr(\mathbf{X}, \mathbf{y} \mid \beta) \Pr(\beta)}{\Pr(\mathbf{X}, \mathbf{y})}.$$

- We can still ignore the "evidence" term $\Pr(\mathbf{X}, \mathbf{y})$ since it is a constant that does not depend on β.
- $\Pr(\beta) = \Pr(\beta_1) \cdot \Pr(\beta_2) \cdot \ldots \cdot \Pr(\beta_d)$
- $\Pr(\beta) \sim$

Bayesian regularization

Easier to work with the **log likelihood**:

$$\arg\max_{\boldsymbol{\beta}} \Pr(\mathbf{X}, \mathbf{y} \mid \boldsymbol{\beta}) \cdot \Pr(\boldsymbol{\beta})$$

$$= \arg\max_{\boldsymbol{\beta}} \prod_{i=1}^{n} e^{-(y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)^2 / 2\sigma^2} \cdot \prod_{i=1}^{n} e^{-(\beta_i)^2 / 2\gamma^2}$$

$$= \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} -(y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)^2 / 2\sigma^2 + \sum_{i=1}^{d} -(\beta_i)^2 / 2\gamma^2$$

$$= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)^2 + \frac{\sigma^2}{\gamma^2} \sum_{i=1}^{d} (\beta_i)^2 / \sigma^2.$$

Choose $\boldsymbol{\beta}$ to minimize $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{\sigma^2}{\gamma^2}\|\boldsymbol{\beta}\|_2^2$.

Completely different justification for ridge regularization!

Test your intuition: What modeling assumption justifies LASSO regularization: $\min \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$?