

Machine Learning

Max-margin linear separator and SVM

Probability, Statistics and Max-Likelihood Principle

DSC 240

Feb 11, 2025

Instructor: Prof. Yu-Xiang Wang

Announcement

- Midterm grading in progress
- HW2 and MP2 due on Thursday.
 - We needed to inspect the autograder a bit longer. Thanks for the patience.
- HW3 will be announced on Thursday. MP3 (final MP) will be announced (hopefully over the weekend).

Recap: Last two lectures

- Linear regression: How to avoid overfitting?
 - Regularization
 - L1 and L2 regularization
- Regularization path: a plot of the coefficients as regularization weight changes.
 - Case study on CA Housing data
- Two flavors of Generalization theory
 - Linear Regression's prediction / estimation error under linear Gaussian model + Fixed Design.
 - "distribution-free" learning bounds on the "Excess Risk" for all learning problems with bounded losses.

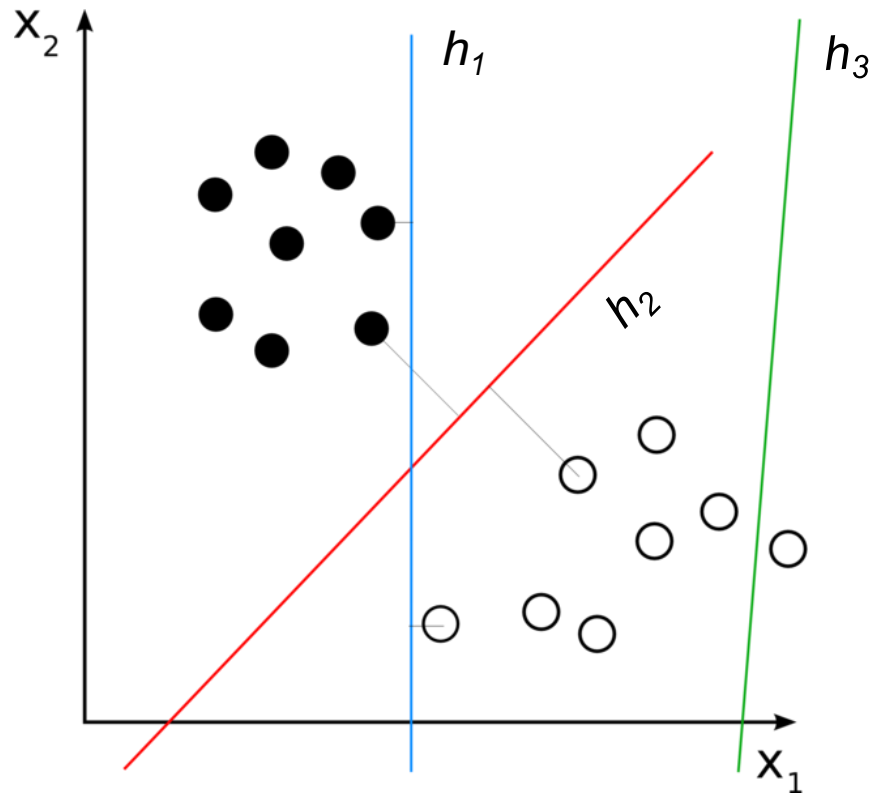
Recap: Regularization can be applied to classification problems too, especially when the number of features is very large.

- L2-regularization: maximize margin
 - Why?
 - Read up on “support vector machine”: Hinge loss + L2-regularization
 - Very interesting geometric insight. It also tells you why we shift “Perceptron loss” to “Hinge loss”
 - I may teach it today if we finish the “midterm review” early.
- L1-regularization: feature selection
 - Promotes a sparse linear combination of the features
 - What’s the benefit?
 - Interpretability
 - Good generalization despite a larger number of irrelevant features.

Today

- Max-margin linear separator and Support Vector Machines
- Probability / Statistics
 - Probability review (I will mostly skip through)
 - Statistics and the statistical estimation problem
 - Maximum likelihood estimation
- Learning goals: Principled derivation of Hinge loss, logistic loss and square losses...

Recap: Is h_2 better than h_1 ?



Which is better, h_1 , h_2 , h_3 ? Why?

Optimization basics in one slide

- Key concepts:
 - Objective function / Criterion
 - Constraints
 - Variables / Arguments
 - Optimal (objective) value
 - Optimal solution
- Example: “Running a Convenient Store”
 - f : Business cost minus revenue
 - θ : Inventory choices
 - Θ : Size of the storage space, available fund.

$$\min_{\theta \in \Theta} f(\theta)$$

Why L2-regularization is associated with maximizing geometric margin?

- How is the “geometric margin” defined?
- Write down the optimization problem for maximizing margin while perfectly classifying points that are linearly separable

Deriving an equivalent, but computationally more convenient, optimization problem

- Original problem:
$$\max_{\mathbf{w}, b} \frac{1}{\|\mathbf{w}\|_2} \min_{i \in [n]} |\mathbf{w}^T \mathbf{x}_i + b| \quad s.t. \quad \forall i \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 0$$
- Observation 1: Does the scale of w and b matter?
- Observation 2: What are ways to change the criterion without changing the solution?
- Observation 3: What happens if we add $\min_{i \in [n]} |\mathbf{w}^T \mathbf{x}_i + b| = 1$ to the constraints?

Hard Margin Support Vector Machines

$$\min_{\mathbf{w}, b} \|\mathbf{w}\|^2$$

$$s.t. \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 \quad \forall i \in [n]$$

- This is a “Quadratic program”.
- Can be solved using any of the off-the-shelf optimization tools, e.g., CPLEX.

Hinge Loss and Soft-Margin Support Vector Machines (with L2-regularization)

$$\min_{\mathbf{w}, b} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i$$

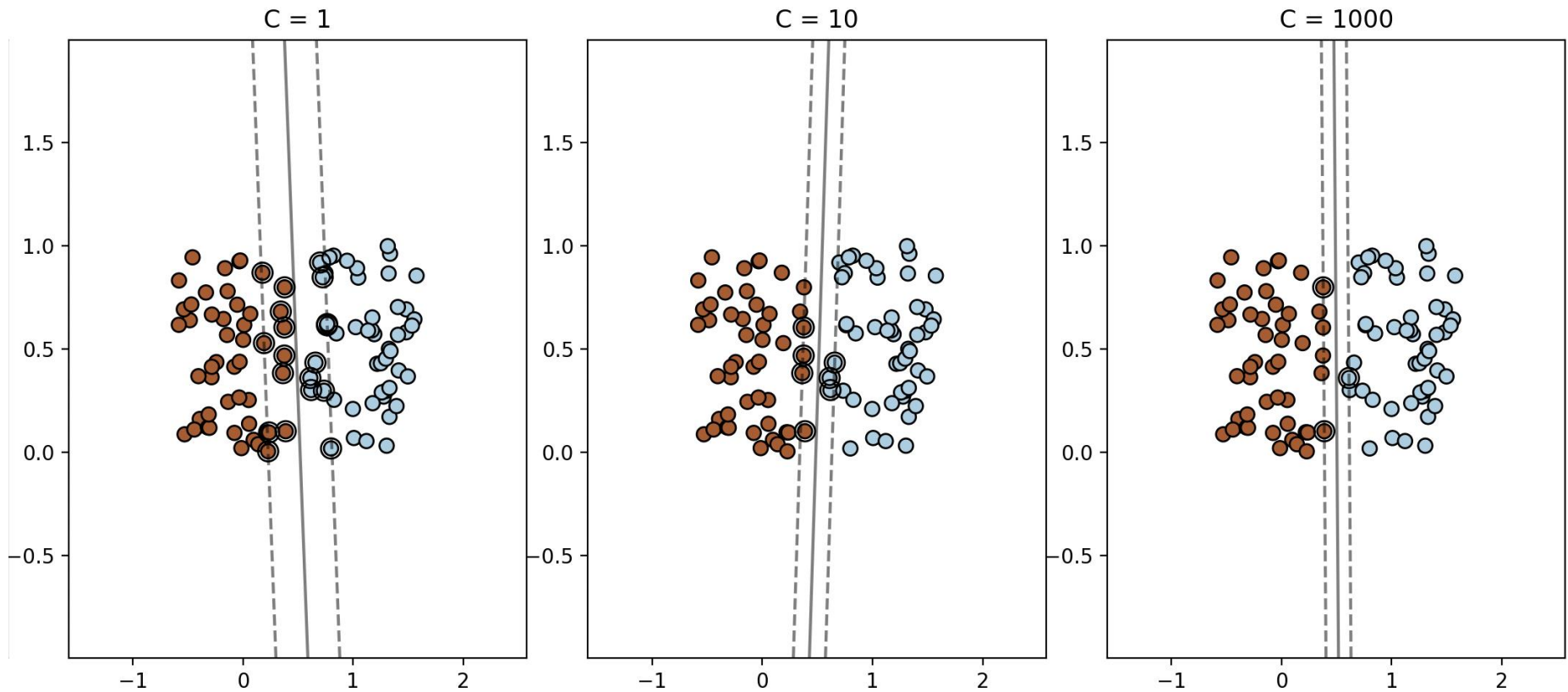
$$s.t. \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i \quad \forall i \in [n]$$

$$\xi_i \geq 0 \quad \forall i \in [n]$$

Equivalent to minimizing **Hinge losses**.

$$\min_{\mathbf{w}, b} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \max [1 - y_i(\mathbf{w}^T \mathbf{x} + b), 0]$$

Choice of hyperparameter C in **soft-margin SVMs** and how they affect the margins and “support vectors”.



Checkpoint: Hinge loss and SVMs

- We derived where hinge loss is coming from
- And why L2-regularization is associated with maximizing margin in binary classification

So far we have learned most elements of machine learning, but...

- We learned to specify a hypothesis class
- We learned to work out the possible shapes of decision boundaries
- We learned how to “train” an ML model --- by solving an optimization problem
- But how did we come up with the hypothesis classes in the first place?
 - We brainstormed... and concluded that (1) decision trees (2) linear-classifiers --- i.e., thresholding a weighted linear combination of features.
 - But how do we know the resulting decision boundaries are appropriate for the problems we hope to solve?

We learned about directly modelling the predictive functions. There is another way... called “Probabilistic modelling”

- We can model how the data is generated in the first place.
 - Model the labeling process via a **conditional distribution** $P(y | x)$. This is known as a ***(probabilistic) discriminative model***.
 - Specifying decision-trees / linear classifiers / shapes of decision boundaries should be considered non-probabilistic discriminative models.
 - Model the **joint distribution** $P(x,y)$. Often one models the label distribution $P(y)$ and a generative process $P(x|y)$. This is known as a ***generative model***.
- The natural prediction would be
 - $h(x) = \operatorname{argmax}_y P(y | x)$
 - If the data generative process is indeed $P(y|x)$, then this is “Bayes optimal”.

What's the benefit of a probabilistic models

- Quantifying uncertainty
 - Soft prediction $P(y|x)$. Not just $\text{argmax}_y P(y|x)$
 - More informative than the score function
- Easy to impute missing features (if we have the joint distribution $P(y,x)$ --- simply marginalize out the missing data)
- More interpretable / explainable. Modeling assumption more easily testable.
- Can do many other things besides classification...
 - Example: Language model / ChatGPT

To learn how ChatGPT works, we need to learn probabilities!

Concepts in Probability

- Probability
- Random variables
- Independence
- Expectation
- Conditioning

Random variable and distribution

- A **random variable** X is a numerical outcome of a random experiment
- The **distribution** of a random variable is the collection of possible outcomes along with their probabilities:
 - Discrete case: $p(X = x) = p(x)$
 - Continuous case: $p(a \leq X \leq b) = \int_a^b p(x)dx$

Expectation

- For a random variable $X \sim p(X = x)$, its **expectation** is

$$\mathbb{E}(X) = \sum_x xp(X = x)$$

- In an empirical sample, x_1, x_2, \dots, x_N , $\mathbb{E}[X] = \frac{1}{N} \sum_{i=1}^N x_i$
- Continuous case: $E[X] = \int_{-\infty}^{\infty} xp(x)dx$
- Expectation of sum of random variables

$$\mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2]$$

- How about the expectation of $f(X)$ --- a function of random variable?

Variance

The **variance** of $f(x)$ denoted as $\text{var}[f]$ provides a measure of how much variability there is in $f(x)$ around its mean value $\mathbb{E}[f(x)]$.

$$\text{var}[f] = \mathbb{E} [(f(x) - \mathbb{E}[f(x)])^2] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

Joint distribution, conditional distribution and marginal distributions

- Probability distribution of many (possibly dependent) random variables.
- Joint: $P(X,Y)$
- Conditionals: $P(X|Y), P(Y|X)$
- Marginals: $P(X), P(Y)$

(Statistical) Independence

- Not the same as linear independence in linear algebra!
- X and Y are independent, i.e.,

$$X \perp Y \text{ iff } P(X, Y) = P(X)P(Y) \text{ iff } P(X) = P(X|Y)$$

- X and Y are independent implies

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$$

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$$

Notations from probability

$\mathbb{E}_{\mathcal{D}}$ [Function of an r.v. X]

$\mathbb{P}_{\mathcal{D}}$ [Event]

$f_{X \sim \mathcal{D}}(x)$ Short hand: $P(X)$

$F_{X \sim \mathcal{D}}(x)$

Conditional expectation / conditional probability / density

$\mathbb{E}[\text{Func}(X, Y)|Y]$

$\mathbb{P}[\text{Event_of}(X, Y)|Y]$

$f(x|y)$

Quiz: Let X, Y be random variables.
What is random and what not?

$$\|X\|^2$$

$$f_{X \sim \mathcal{D}}(x)$$

$$\mathbb{E}[f(X)]$$

$$F_{X \sim \mathcal{D}}(x)$$

$$\mathbb{P}[X > 5]$$

$$\mathbb{P}[X > 5 | Y^2 < 18]$$

$$\mathbb{P}[X > 5 | Y]$$

$$\mathbb{E}[X | Y = y]$$

$$\mathbb{E}[\det(X) | Y]$$

Statistics in one slide

- What is the difference between probability and statistics?
 - Statistics is the “science of data” --- it uses probability theory, but also other branches of mathematics and computational tools for making sense of data
- Typical problem that you learn in a first course of statistics --- Statistical estimation
 - Data: $X_1, X_2, \dots, X_n \sim P$
 - Goal: Estimate a statistical quantity θ of the distribution P (not a function of the data!)
 - **Estimator** (really an algorithm): $\hat{\theta}$ that takes input data and output an guess of the true quantity θ
- Examples
 - Estimating the mean, variance, medians (and other quantiles)
 - Estimate the expected error of a given classifier using a holdout dataset.
 - Estimate the parameter θ of P if P is parameterized by θ , denoted by P_θ .

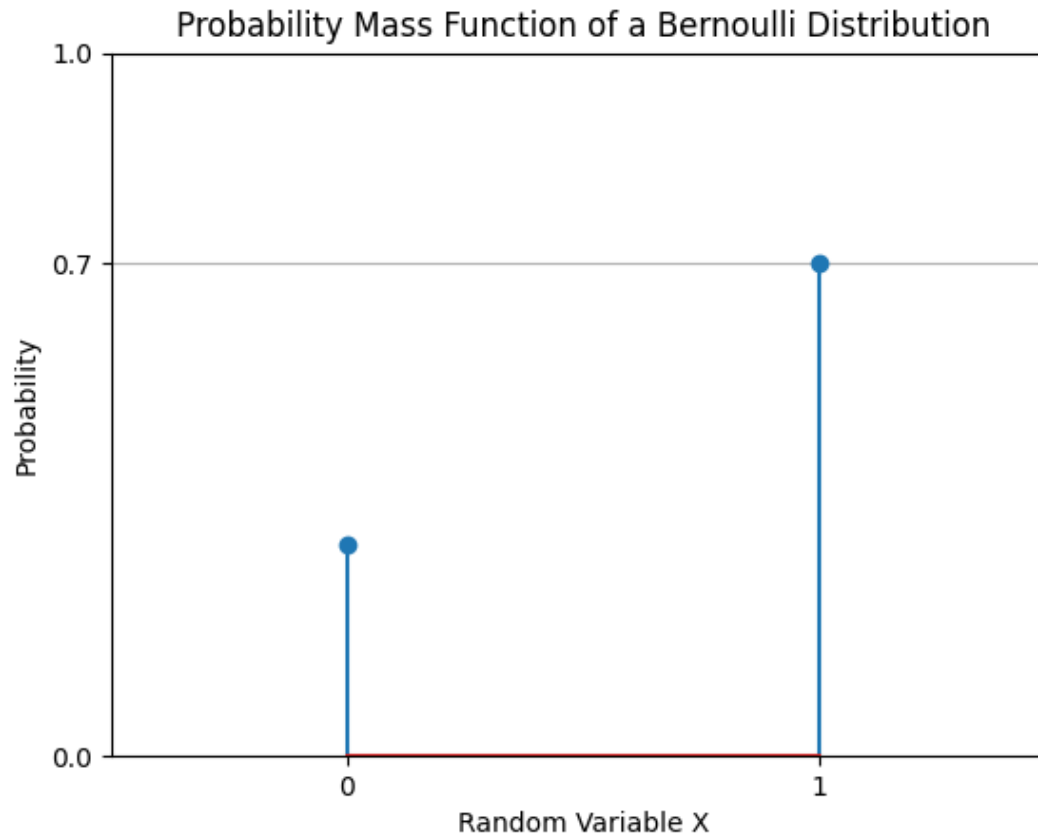
Examples of statistical estimation problem

- Example 1 (Biased Coin): Toss a coin 100 times, observe the outcome {Head or Tail}. What is the probability of seeing “Head”?
 - Application to ML. How to estimate classification error on future data?
- Example 2 (Average Monthly Precipitation in La Jolla):
 - Observe data for Year 1960, 1961,.....,2023.
 - Each data point is a vector of 12 numbers measuring the number of inches of precipitation.
 - How to estimate the average?

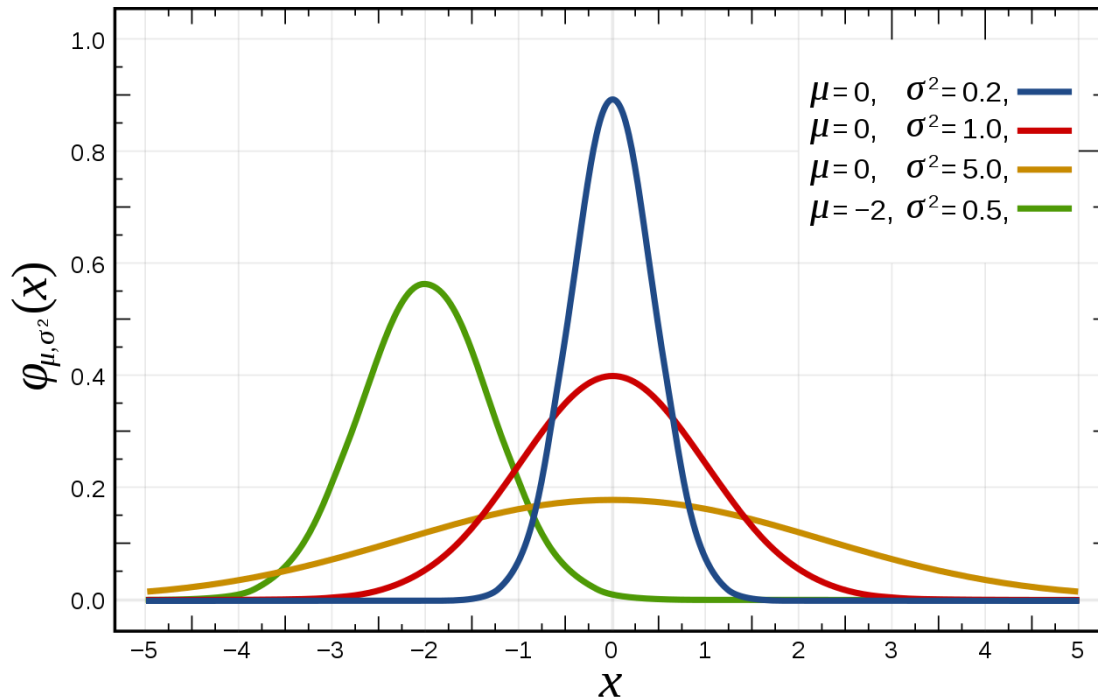
Bernoulli Distribution

$$X \sim \text{Ber}(p)$$

$$P(X = x) = \begin{cases} p & \text{if } x = 1, \\ 1 - p & \text{if } x = 0. \end{cases}$$



Gaussian distribution $X \sim \mathcal{N}(\mu, \sigma^2)$



(figure from wikipedia)

- Probability density: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$

Maximum likelihood estimation

- Used since Gauss, Laplace, etc.... Popularized / carefully analyzed by Ronald Fisher.
- Which distribution is more *likely* to have produced the data?

$$\max_{P \in \Pi} f_{\text{Data} \sim P}(\text{Data})$$

- Observation 1: If the data is i.i.d. the by independence the density factorizes
- Observation 2: Taking log does not change the solution.

Exercise: Estimating Bernoulli Mean using MLE

- Data: $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} \text{Bernoulli}(p)$
- Likelihood: $\mathcal{L}(X_i; p) = p^{X_i} (1 - p)^{1 - X_i}$
- The MLE problem: $\hat{p} = \arg \max_{p \in [0, 1]} p^{\sum_i X_i} (1 - p)^{n - \sum_i X_i}$

Exercise: Estimating the mean parameter of a Gaussian distribution

• Data $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$

• Likelihood: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{x - \mu}{\sigma}\right)^2\right)$

• The MLE problem: $\hat{\mu} = \arg \max_{\mu \in [0,1]} \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X_i - \mu)^2}{2\sigma^2}}$

Exercise: Linear regression

- $P(y|x)$ is modeled by “Linear Gaussian model”

$$y_i = x_i^T \theta^* + \epsilon_i \quad \text{where } \epsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}$$

- Data: $(x_1, y_1), \dots, (x_n, y_n)$
- Work out the optimization problem to solve for the MLE for θ^* .

Exercise: Logistic regression

- $P(y|x)$ is modeled by a Logit model $y \sim \text{Bernoulli}(\text{Sigmoid}(x^T \theta^*))$

$$\text{where Sigmoid}(t) = \frac{e^t}{e^t + 1}$$

- Data: $(x_1, y_1), \dots, (x_n, y_n)$
- Work out the optimization problem to solve for the MLE for θ^* .

Checkpoint

- We learned to derive SVM (hinge loss + l_2 -regularization) by a geometric “max-margin” principle.
- We learned about probabilistic models and how to derive other loss functions using what we learned.
- Next lecture: Naïve-bayes classifier and other generative models.